De la recherche d'information orientée sytème à la recherche d'information orientée contexte : Verrous, contributions et perspectives

Lynda Tamine

Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse CEDEX 9
IRIT- Institut de Recherche en Informatique de Toulouse, UMR 5505

Mémoire de Synthèse

présenté en vue de l'obtention de

l'Habilitation à Diriger des Recherches

Spécialité : Informatique

par

Lynda TAMINE-LECHANI

De la recherche d'information orientée système vers la recherche d'information orientée contexte : Verrous, contributions et perspectives

Soutenue le 25 Novembre 2008 devant la commission d'examen :

M.	C. Chrisment	Professeur	Université de Toulouse III	(Examinateur)
Mme	M. Beaulieu	Professeur	Sheffield University	(Rapporteure)
M.	M. Bouzeghoub	Professeur	Université de Versailles	(Rapporteur)
Mme	M.C. Fauvet	Professeur	Université de Grenoble 1	(Rapporteure)
Mme	C. Berrut	Professeur	Université de Grenoble 1	(Examinatrice)
M.	M. Boughanem	Professeur	Université de Toulouse III	(Directeur de Recherches)

à Katia et Liza,
à Mourad,
à mes parents.

Remerciements

Je tiens tout d'abord à exprimer mes vifs remerciements à l'ensemble des rapporteurs de ce travail :
Mme Micheline Beaulieu, Professeur à l'Université de Sheffield, Mme Marie-Christine Fauvet, Professeur à l'Université de Grenoble I, M. Mokrane Bouzeghoub, Professeur à l'Université de Versailles. Leurs remarques pertinentes et constructives m'ont permis d'améliorer la qualité de ce manuscrit, d'avoir des avis éclairés sur le travail accompli et sur les directions de recherche futures.
Je remercie également Mme Catherine Berrut d'avoir accepté d'examiner mon travail et faire partie du jury.

Je tiens à remercier vivement, peut être jamais assez, M. Mohand Boughanem qui dirige mes travaux de recherche depuis mon Magister. Je le remercie très sincèrement pour tous les moyens qu'il a mis à ma disposition pour mener à bien mes travaux, pour la confiance qu'il a toujours témoignée à mon égard. Je saisis cette occasion, pour dire combien j'apprécie, après treize années de collaboration, sa compétence, et son sens de la persévérance.
Je me permets, au nom de notre composante, de le saluer pour les gros efforts qu'il fournit pour mettre en place un cadre de travail propice à nos activités de recherche.

Mes vifs remerciements vont également à M. Claude Chrisment, qui m'a accueillie dans son équipe, qui m'a prodigué son soutien indéfectible pour mes activités de recherche et d'enseignement. Je le remercie très sincèrement pour ses conseils, ses avis éclairés et ses encouragements sans cesse renouvelés.

La synthèse des travaux de recherche présentés dans ce document est le résultat d'un travail d'équipe réalisé en partie avec les thèsards, principalement Fatiha Boubekeur et Mariam Daoud. Je tiens donc à les remercier avec intensité pour le sérieux, la curosité et la grande capacité de travail dont elles ont fait preuve. Je remercie également, Ba Dinh, Ourdia Boudghaghèn, Samir Kechid et Amira Tifous pour leur grande volonté de travail.

Merci à tous les membres de l'équipe SIG. Un grand merci à Wahiba, Karen et Cécile pour leur convivialité. Les discussions que nous menons au quotidien sur des sujets très divers ont donné une ambiance très agréable au déroulement de mes travaux. Encore merci !

Mes vifs remerciements et non des moindres, vont à l'égard de Mourad, Liza et Katia. Je remercie particulièrement Mourad pour son soutien indéfectible depuis de longues années. Je ne le remercierai jamais assez pour sa compassion, ses encouragements, sa disponibilité et ses conseils. Je saisis cette occasion pour exprimer des choses que je ne sais pas trop exprimer au quotidien.

Enfin, je remercie avec intensité mes parents pour leur soutien. Un grand merci également à mes frères et soeur, ma belle-famille pour leur aide et leur disponibilité.

Table des matières

Liste des figures

1 De la RI orientée système vers la RI orientée contexte : génèse et positionnement de nos travaux

1.1 Introduction

Les travaux présentés dans ce document s'inscrivent dans le cadre de mes activités de recherche au sein de l'équipe Systèmes d'Information Généralisés (SIG), plus précisément, de la composante Recherche d'Information (RI) du laboratoire CNRS/IRIT de l'Université Paul Sabatier. Leur cadre général porte sur la RI, domaine déjà ancien (début des années 60), qui n'a cessé d'évoluer dans le but de rationaliser le processus complexe permettant l'identification, au sein de volumes de plus en plus importants d'informations, celles qui sont potentiellement intéressantes pour l'utilisateur. Cette évolution a été tout d'abord marquée par l'ère, plus connue sous le nom de *RI orientée système* qui a mis l'accent sur les technologies permettant l'amélioration des fonctionnalités de base des Systèmes de Recherche d'Information (SRI) indépendamment de leurs utilisateurs. La RI orientée système a permis la mise en place d'approches formelles et modèles théoriques permettant l'accès à l'information, assujetties à des méthodologies d'évaluation de leur efficacité. Les résultats tangibles issus de cette approche se traduisent par la mise en œuvre de la première génération de SRI capables d'indexer des volumes considérables d'information et de les apparier avec des requêtes utilisateurs, sous des conditions d'efficacité acceptables et contrôlables. Les premiers acquis scientifiques du domaine, à travers le courant de la RI orientée-système, constituent incontestablement les fondements de l'ensemble des évolutions qui s'en sont suivies. Un des courants de recherche ayant marqué l'ère suivante, est véhiculé par la vision de la *RI orientée utilisateur*, apparue en 1977, qui a essentiellement porté sur les aspects liés aux interactions entre un utilisateur et un système d'accès à l'information et ce, à travers les phases d'expression des besoins, de perception de l'information et de la définition de la pertinence à des fins d'adaptation. Ce courant a été entretenu, voire ravivé, par l'essor du *web* qui a remis l'utilisateur au centre du processus de recherche d'information. Une vision plus large, connue sous le le nom de *RI cognitive*, a alors lancé des défis liés à l'interprétation du besoin en information

dans le cadre d'une tâche ou d'une situation, de l'interdépendance des éléments de l'environnement de l'utilisateur et leur impact sur sa perception de la pertinence (Belkin 1978). Cette vision est résumée par les fondements énoncés par (De Mey 1977) :

1. *information processing takes place in senders and recipients of messages,*
2. *processing takes place at different levels,*
3. *during communication of information, any actor (either sender or recipient) is influenced by its past and present experiences (time) and its social, organizational and cultural environment,*
4. *individual actors influence the environment or domain,*
5. *information is situational and contextual.*

Ces fondements, qui ne disent rien des méthodes à employer, ni des verrous à lever, pointent avec précision l'objectif : contextualiser le traitement de l'information. En clair, le problème n'est pas tant la disponibilité de l'information mais son accessibilité relativement à un contexte d'utilisation particulier. Les besoins sont ainsi énormes.

Nos travaux s'inscrivent précisément dans le courant de cette évolution de la RI orientée système vers la RI contextuelle, visant l'adaptation du processus de recherche d'information pour des utilisateurs spécifiques. Cette spécificité porte d'une part sur l'utilisateur, qui est au centre de l'activité de recherche d'information et d'autre part sur le contexte de cette activité, caractérisé par des dimensions relevant de choix que nous avons effectués.
Nos contributions portent sur deux principaux volets. Le premier volet concerne la spécification et la formalisation d'un modèle adaptatif/contextuel (marqué par l'évolution de nos objectifs) d'accès à l'information, plus précisément de type texte. Ces travaux se déclinent par des investigations diverses, focalisées sur la clarification et formalisation des besoins en information, de la modélisation du contexte de recherche puis de son intégration dans le modèle d'accès à l'information. Ensuite, comme un modèle n'est viable que lorsqu'il est reconnu efficace selon des normes et méthodologies d'évaluation reconnues, nous nous sommes intéressés, dans un second volet, à la définition d'un cadre d'évaluation permettant la validation de nos contributions dans le domaine.
Ce chapitre d'introduction est organisé comme suit. La section 1.2 donne une brève rétrospective des réflexions menées par la communauté dans le sens de l'adaptation du processus de recherche d'information à l'utilisateur. La section 1.3 en développe les perspectives vers une vision plus large de l'utilisateur dans son environnement de recherche, véhiculée par la RI contextuelle. La section 1.4 présente les verrous scientifiques et technologiques posés par la RI contextuelle. La section 1.5 met l'accent sur nos investigations de recherche. Nous y rapportons notamment nos motivations, cernons les verrous que nous avons explicitement ciblés durant nos activités de recherche et donnons un aperçu de nos contributions dans le domaine. La section 1.6 annonce le plan du document.

1.2 La RI orientée système vs. la RI orientée utilisateur

La RI orientée-système est l'approche pionnière dans le domaine, apparue dès les années 1960. L'objectif majeur des travaux issus de cette approche était de faire asseoir des modèles, des techniques et des algorithmes permettant d'identifier, dans un corpus de documents, les informations (documents ou parties de documents) pertinentes en réponse à un besoin en information exprimé par l'utilisateur à travers une requête. La pertinence est une valeur de jugement dichotomique de l'information (pertinente/non pertinente) appréhendée selon l'aspect thématique en ce sens que l'information est qualifiée de pertinente dans le cas où elle porte sur le sujet de la requête, elle est qualifiée de non pertinente dans le cas contraire. L'identification ou sélection de l'information pertinente est une activité qui est impliquée par divers modèles et algorithmes qui reposent sur un processus coordonnant différentes étapes :

1. la formulation de la requête via un langage d'expression peu voire pas formel,
2. la représentation des documents à l'aide de structures décrivant leur contenu,
3. la réécriture de la requête pour des besoins d'optimisation et/ou de formalisation,
4. l'évaluation de la requête par appariement avec l'information contenue dans le corpus,
5. présentation des résultats de l'évaluation sous forme d'informations supposées potentiellement pertinentes par le système.

Les investigations menées dans le cadre de l'approche orientée-système ont essentiellement porté sur la spécification de représentations de documents et de requêtes, la définition d'algorithmes d'appariement entre ces représentations et la mise en œuvre de méthodologies d'évaluation de l'efficacité de ces modèles (Baeza-Yates et Ribeiro-Neto 1999). Le paradigme de l'évaluation est, par essence, fondé sur une comparaison de la pertinence thématique annoncée par les assesseurs[1] et pertinence dite algorithmique, estimée par le SRI.
Les résultats phares obtenus durant la période $1960 - 1990$ sont de façon non exhaustive :

- le développement de modèles mathématiques de RI : modèle vectoriel (Salton 1968) et modèle probabiliste durant la période 1960 à 1970 (Robertson 1977) puis modèles logiques durant les années 1980 (van Rijsbergen 1986a),

- le développement de méthodes de classification et de catégorisation de textes durant la période 1960 à 1970 (Beorko et Bernick 1963),

[1] Personnes participant aux campagnes d'évaluation chargées de juger manuellement la pertinence des informations issues de l'évaluation de requête

- le développement de stratégies de RI adaptative : réinjection de la pertinence et d'expansion de requête (Rocchio 1971), durant les années 1970,

- l'application de techniques d'analyse du langage naturel durant les années 1980 (Church et al. 1989),

- le développement d'une méthodologie d'évaluation en RI basée sur le paradigme de Cranfield durant les années 1960 (Cleverdon 1967a).

Le début des années 1990 a marqué une période de profusion des travaux de recherche dans le domaine de la RI de manière générale. Cet essor a été encouragé d'une part par l'apparition de TREC[2] (Harman 1992b), et d'autre part, motivé par la généralisation du *web* qui a posé de nouvelles problématiques telles que la recherche d'information à grande échelle, la recherche d'information précise, la détection de *spam* etc. Notre objectif n'étant pas de focaliser sur les contributions de la RI orientée-système à l'état de l'art du domaine, nous invitons le lecteur intéressé à se reporter aux synthèses publiées dans (Allen 1991, Kantor 1994).

Parallèlement à l'approche orientée-système, plus précisément en 1977, est né un courant de recherche, matérialisé par l'approche utilisateur et approche cognitive de la RI, mettant l'utilisateur au centre de l'activité de recherche d'information (Brookes 1977, Järvelin 1986, Ingwersen 1996b). Soulignons de prime abord que la différence entre les deux approches est peu perceptible (Ingwersen et Jarvelin 2005), fondée sur l'accent que met l'approche orientée utilisateur à exploiter la dimension cognitive dans le contexte de la RI sur le web spécifiquement.
De nombreuses études critiques (Dervin et Nilan 1986, Shamber 1994, Ingwersen 1996a) reconnaissant la plus value indéniable de la RI orientée-système, contestent cependant la non-considération de la dimension utilisateur dans le processus de base de recherche d'information, qui, par essence, est initié par l'utilisateur et au final produit des résultats qui lui sont destinés. Des travaux ultérieurs s'intéressent alors à l'intégration de la perspective cognitive de l'utilisateur (Brookes 1977, Järvelin 1986, Ingwersen 1996b) dans l'interprétation du concept information (Belkin 1978) et dans l'interprétation du besoin en information dans le cadre d'une tâche ou d'une situation ; ces travaux s'intéressent également à l'appréciation de l'interdépendance des éléments de l'environnement de l'utilisateur et leur impact sur sa perception de la pertinence (Belkin 1987). Sans être antagoniste à la RI orientée-système, la RI orientée utilisateur est à l'origine de trois révolutions nommées explicitement dans (Robertson et Hancock-Beaulieu 1992) :

[2]Text REtrieval Conference

1. *la révolution cognitive* : cette révolution remet en cause l'isolation du besoin en information de l'acteur l'ayant exprimé. En ce sens que l'interprétation du besoin en information est subordonné à la dimension cognitive de l'utilisateur qui l'a exprimé (centres d'intérêt, expertise, objectifs, tâche etc.) et de la situation de recherche correspondante,

2. *la révolution de la notion de pertinence* : cette révolution remet en cause l'aspect statique et dichotomique de la pertinence, qui est le concept clé en RI. La pertinence est, sous l'angle de la RI orientée utilisateur, un concept multidimensionnel couvrant divers niveaux (très pertinent, peu pertinent, marginalement pertinent) et types (pertinence situationnelle, pertinence affective, pertinence cognitive) (Borlund 2003a;b),

3. *la révolution de l'interaction* : cette révolution remet en cause la séparation entre l'utilisateur et le SRI en ce sens que l'interaction utilisateur-SRI est une dimension non négligeable, devant être clairement intégrée dans la conception globale du processus de recherche d'information.

Les premiers résultats fondamentaux issus de la RI cognitive/utilisateur sont notamment :

- le développement de modèles cognitifs de l'interaction en RI (Saracevic et al. 1991, Vakkari 2001),

- le développement de méthodes pour la multi-représentation de documents et requêtes représentant les différentes visions cognitives de leur contenu (Ingwersen 1994b, Larsen et al. 2003, Tamine et al. 2003),

- le développement de modèles et stratégies de filtrage d'information (Belkin et Croft 1992) et de recommandation (Lieberman 1995, Pazzani et al. 1996)

- le développement de modèles et de stratégies d'accès contextuel/personnalisé à l'information (Belkin et al. 1996, Mc Gowan 2003, Micarelli et al. 2007)

- le développement de modèles d'évaluation basé sur la pertinence situationnelle dans le contexte d'une recherche d'information interactive (Robertson et al. 1992).

Le tableau 1.1 résume nos propos en mettant en exergue les différences d'interprétation et/ou perception des concepts fondamentaux en RI (information, besoin en information, utilisateur, interaction etc.) selon la vision orientée-système vs. vision orientée-utilisateur.

	RI orientée-système	**RI cognitive/orientée-utilisateur**
Information	Termes isolés représentant des facteurs d'indexation et/ou appariement	Résultat de l'interprétation des mots en relation avec la vision cognitive de l'utilisateur
Besoin en information	Mots-clés isolés de l'utilisateur	Mots-clés intégrés à la vision cognitive de l'utilisateur, de sa tâche, sa situation ...
Utilisateur	Acteur exprimant en information en dehors des frontières du processus de recherche d'information	Acteur intégré au processus de recherche d'information
SRI	Retourne les mêmes résultats en réponse aux mêmes requêtes	retourne les mêmes résultats en réponse aux mêmes requêtes émises dans les même situations de recherche (utilisateur, tâche, ...)
Modèle de RI	Spécification de modèles de représentation de requêtes, documents et d'appariement requête-document	Le modèles est placé dans le contexte cognitif de l'ensemble des acteurs (utilisateur, auteur, concepteur du SRI)
Pertinence	Thématique, algorithmique	Situationnelle, affective, cognitive
Efficacité	Capacité du SRI à identifier des documents pertinents du point de vue thématique à des requêtes thématiques (perçues comme des sujets de besoins en information)	Capacité du SRI à satisfaire le besoin en information d'un utilisateur spécifique dans un contexte de recherche spécifique

TAB. 1.1 – *RI orientée système vs. RI orientée utilisateur/cognitive*

Demeurant dans la perspective de l'adaptation du processus de recherche d'information à l'utilisateur et son environnement, l'un des résultats phares issus de la vision utilisateur (cité ci-avant) et qui a constitué le contour d'une grande partie de nos travaux, est la RI contextuelle. C'est une direction de recherches qui vise particulièrement la prise en compte du contexte de recherche, comme une nouvelle dimension, permettant à terme de déployer un appariement requête-contexte-document.
La notion de *contexte* trouve son origine et historique dans de nombreuses disciplines (Schilit et al. 1994, Ryan et al. 1997, Davies et al. 1998) telles que la modélisation utilisateur, l'intelligence artificielle, l'hypermédia adaptatif, l'informatique *ubiquitaire*, les bases de données (Ioannidis et Koutrika 2005, Kostadinov et al. 2008). C'est une notion difficilement formalisable, ayant de multiples définitions couvrant les différentes facettes qu'elle couvre, selon le cadre d'application associé

(Bazire et Brézillon 2005).

La section suivante apporte un éclairage sur la notion de contexte dans le cadre particulier de la RI, sur les problématiques et origines de la RI contextuelle ainsi que sur les spécificités d'un SRI contextuel. Cette présentation nous permettra de mieux cerner les verrous posés dans un tel cadre, ainsi que les solutions apportées dans la communauté.

1.3 Vers la RI contextuelle

1.3.1 Notion de contexte en RI

Les premiers travaux de Saracevic (Saracevic 1997) définissent le contexte selon un modèle cognitif par lequel on peut identifier des structures ou espaces cognitifs qui sont autant de variables impliquées dans le processus de RI et qui peuvent décrire les intentions et les perceptions de l'utilisateur de ce qui l'entoure. Ces variables sont l'espace cognitif de l'utilisateur, l'environnement social ou organisationnel, les intentions et les buts de l'utilisateur ainsi que le système lui-même. Par la suite, de nombreux travaux ont identifié des niveaux de description du contexte selon les familles de facteurs caractéristiques tels que le niveau environnement, le niveau utilisateur et le niveau interaction (Cool et Spink 2002) ou alors selon un modèle basé sur des infrastructures incluant de larges classes contextuelles liées à la tâche, à l'utilisateur, au système etc. (Ingwersen 1996a).

Notre revue de la littérature dans ce domaine nous a amenés à synthétiser puis définir une taxonomie des dimensions du contexte en RI, présenté dans la figure 1.1. Nous y distinguons cinq (5) principales dimensions abordées ou explorées en RI contextuelle, que nous présentons ci-après en explicitant les facteurs associés ainsi que la motivation de l'adaptation du processus de recherche à chacun d'eux, traduisant ce qui est communément appelé *contextualisation de l'accès à l'information*.

1. **Moyen d'accès à l'information** : représente l'outil physique permettant d'effectuer un accès direct à l'information tel que l'ordinateur, le téléphone portable, le PDA[3] etc. L'adaptation du processus de recherche d'information aux caractéristiques de l'outil physique d'accès s'est imposé particulièrement pour des utilisateurs mobiles présentant des contraintes d'ordre situationnelle (requêtes courtes) et physiques (ressources mémoires limitées, zone d'affichage des résultats réduite) (Göker et Myrhaug 2002).
2. **Contexte spatio-temporel :** cette dimension comprend deux sous-dimensions portant sur la localisation géographique et le temps. L'adaptation selon cette dimension concerne particulièrement les applications où les informations ont une validité subordonnée au lieu et temps instanciés lors de l'activité de recherche d'information (routage, guide touristique etc.) (Tao et al. 2003, Göker et Myrhaug 2008).

[3] *Personal Digital Assistant*

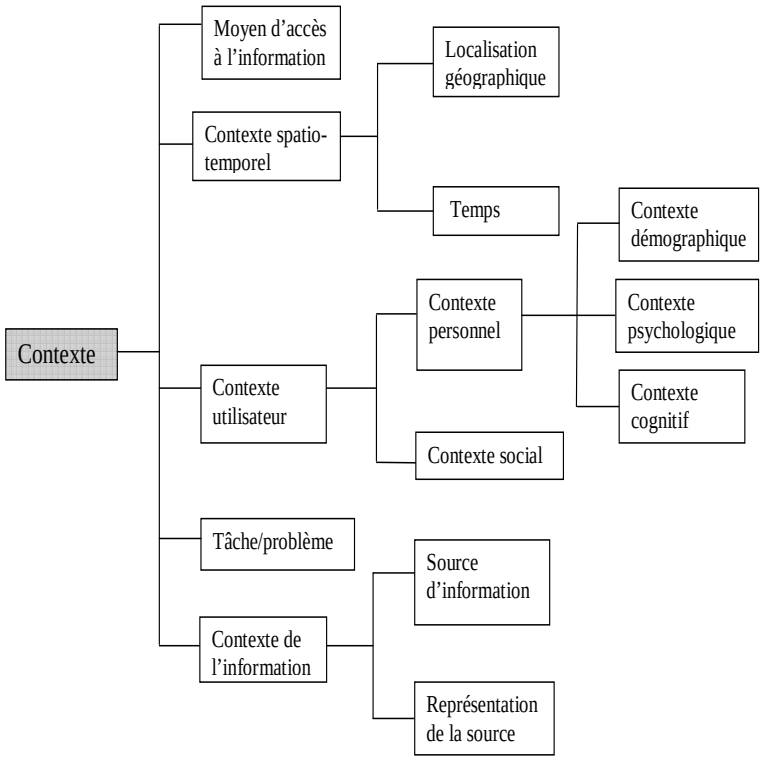


FIG. 1.1 – *Le contexte : une notion multidimensionnelle*

3. **Contexte utilisateur :** c'est la dimension principale abordée par la communauté. Cette dimension comprend deux sous-dimensions : contexte personnel et contexte social.
   (a) *Contexte personnel* : comprend à son tour les sous-dimensions suivantes :

   – ***Contexte démographique*** : porte sur des facteurs de préférences personnelles tels que la langue (*Google personalized, Yahoo*) et le *sexe* (Hupfer et Detlor 2006, Frias-Martinez et al. 2007), exploitées de manière à personnaliser la recherche pour des besoins spécifiques.

   – ***Contexte psychologique*** : *l'anxiété* et *la frustration* sont des exemples de facteurs ayant un impact sur le comportement de l'utilisateur notamment son jugement de pertinence (Bilal 2000, Kim 2008).

   – ***Contexte cognitif*** : cette sous-dimension est sans doute la plus importante. Elle porte particulièrement sur *l'expertise* (Timothy et al. 2005) et les *centres d'intérêt* à court terme

(Agichtein et al. 2006, Haveliwala 2002a, Joachims et al. 2007, Shen et al. 2005, Daoud et al. 2008c) ou centres d'intérêt à long terme de l'utilisateur (Liu et Yu 2004, Sieg et al. 2004a, Tamine et al. 2008a).

(b) *Contexte social* : cette dimension met l'accent sur la *communauté* à laquelle appartient l'utilisateur telle que le amis, les voisins et les collègues. L'adaptation du processus de recherche d'information consiste essentiellement à considérer les préférences et les profils partagés par la communauté de l'utilisateur plutôt que ses préférences et profil personnels (Lang 1995, Smyth et Balfe 2006).

4. **Tâche/problème :** cette dimension porte sur l'intention de l'utilisateur induite par l'expression de sa requête. L'analyse de cette dimension a permis de définir des classes d'intention telles que la *recherche ciblée* (*fact-finding*) vs. *recherche exploratoire* (*exploration task*) (Navarro-Prieto et al. 2006) ou alors la tâche transactionnelle, la tâche informationnelle ou la tâche navigationnelle, évoquant particulièrement une recherche d'information sur le *web* (Jansen et al. 2007). La tâche peut également porter sur la finalité de l'application ou service invoquant une recherche d'information telle que *le guide touristique* (Cheverst et al. 2000) et transport guidé par GPS[4] (Schilit et al. 2003).

5. **Contexte de l'information** : cette dimension trouve son essence dans le principe de polyreprésentation développé dans les travaux de Ingwersen (Ingwersen 1994a) issus de la RI cognitive. Le principe de polyreprésentation de l'information repose sur l'hypothèse que les informations sont classifiables selon un ensemble de critères (considérés comme des dimensions de l'espace de représentation de l'information) tels que : le genre de l'information, son auteur, sa structure, son style etc. et que la considération de l'ensemble de ces critères permet de mieux subordonner la pertinence de l'information à la situation de recherche globale. Cette dimension comporte à son tour deux sous-dimensions. La première porte sur le contexte direct de l'information (forme, couleur, métadonnées de structure etc.) (Tombros et al. 2005). La deuxième source porte sur la source d'information et sa perception par les utilisateurs (Xie 2008).

Nous nous appuyons sur cette taxonomie aussi bien pour rapporter l'état des lieux des connaissances dans le domaine que pour situer l'essence de nos réflexions et délimiter le contour de nos activités de recherche. Avant d'y parvenir, nous illustrons dans ce qui suit, la plateforme générale d'un SRI supportant spécifiquement la notion de contexte puis mettons en évidence sa déclinaison dans le processus d'accès à l'information.

[4]Global Positioning System

1.3.2 Architecture d'un système de RI contextuel

Un SRI est dit *contextuel* ou *sensible au contexte* (*context-aware*) s'il exploite les données du contexte de recherche pour sélectionner l'information pertinente en réponse à une requête utilisateur. Il en ressort que la pertinence de l'information est assujettie à son adéquation à la requête et à l'ensemble ou partie des dimensions du contexte (définies ci-haut) qui sont perceptibles dans la situation de recherche en cours. La figure 1.2 présente l'architecture de base d'un SRI contextuel. On y note particulièrement deux fonctionnalités fondamentales :

1. ***La modélisation du contexte*** : par contraste à la RI orientée-système qui s'appuie sur la requête comme unique source d'évidence à modéliser, permettant de spécifier le besoin en information, la RI contextuelle s'appuie sur une source d'évidence additionnelle exprimée à travers le contexte qu'il convient alors de modéliser. La nature et la portée du modèle dépendent des dimensions du contexte considérées. Le contexte utilisateur étant la dimension la plus abordée dans la littérature, la modélisation du contexte est alors qualifiée souvent de modélisation de l'utilisateur (*user modeling*) (Kobsa et Wahlster 1989, Pohl 1997). De manière générale, un modèle de contexte est défini par instanciation de chacun de ces éléments :
   - *les sources d'information* : environnement (temps, température etc.), collection de documents, historique des interactions etc.
   - *des stratégies de collecte de ces informations* : on distingue principalement entre les stratégies implicites et stratégies explicites pour la collecte des données du contexte,
   - *des ressources de modélisation* : des ressources, généralement sémantiques (ontologies, dictionnaires, ...), sont parfois exploitées pour enrichir les données du modèle,
   - *des modèles de représentation et/ou évolution* : permet de formaliser la représentation du contexte en qualité de structure unifiée (partie d'une ontologie, classe de vecteurs de termes, ensemble de concepts ...) ou d'un ensemble d'informations avec des structures différentes et spécifiques, puis de les faire évoluer en cours du temps.

2. ***L'accès contextuel à l'information*** : c'est le processus classique de RI projeté selon une dimension additionnelle liée au contexte de recherche. Basiquement, son objectif est de sélectionner l'information pertinente à la requête adressée au SRI, tenant compte de la requête d'une part et du contexte de recherche en cours d'autre part.

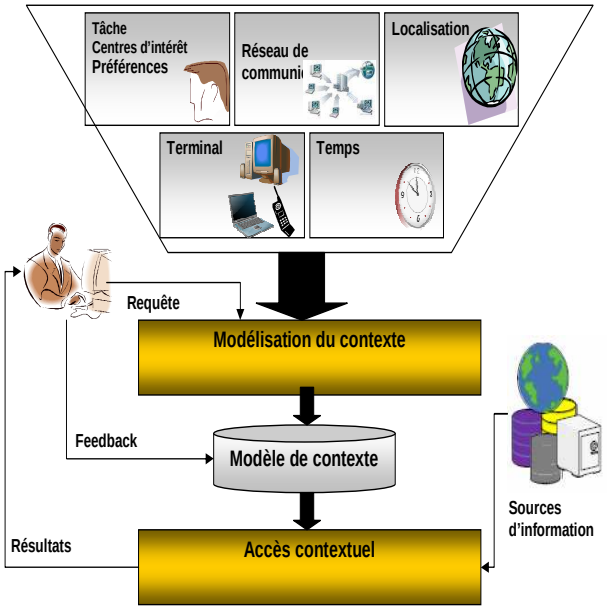


FIG. 1.2 – *Architecture de base d'un SRI contextuel*

1.4 Les verrous scientifiques et technologiques

Cette section a pour objectif, de donner un aperçu des verrous scientifiques et technologiques qui ont marqué et motivé les travaux de recherche visant l'adaptation du processus de recherche d'information pour un contexte/utilisateur spécifique (indifféremment des grandes approches de base système/cognitive) puis de s'en servir comme référentiel pour identifier les verrous que nous avons particulièrement ciblés.
Nous synthétisons dans qui suit les principaux verrous que les travaux du domaine ont tenté de lever.

1.4.1 Formaliser, modéliser, représenter l'utilisateur et son contexte

L'accès contextuel à l'information a pour finalité de délivrer l'information appropriée à l'utilisateur en tenant compte de son contexte. Ce dernier traduit une représentation, généralement partielle, de ce qu'il sait, ce qu'il est, de son lieu, de son expérience, de son groupe etc. Il est clair que la modélisation du contexte est une étape clé dans la conception d'un système d'accès contextuel à l'information (Micarelli et al. 2007). C'est un verrou qui se décline par les questions fondamentales suivantes :

- *Comment identifier les dimensions du profil et de son contexte* : il convient tout d'abord de distinguer les éléments dépendant directement de l'utilisateur de ceux qui dépendent de son environnement ; quelles sont les informations associées ?
- *Quelles sont les interdépendances entre ces dimensions* : recouvrement, causalité, complémentarité ?
- *Quelles sources pour les caractériser* ?
- *Comment formaliser ces dimensions dans le but de les représenter* ?
- *Comment les représenter et les faire évoluer* ?

Ces questions font l'objet d'investigations de recherches dans différentes disciplines : modélisation de l'utilisateur, intelligence artificielle, interaction homme-machine, recherche d'information etc. Dans le domaine précis de la RI, les recherches s'orientent vers la définition de modèles d'utilisateurs permettant à même de préciser les résultats d'une activité de recherche d'information. Les questions précédentes restent posées, les trois dernières s'affinent précisément vers : (1) l'identification de sources d'informations appropriées traduisant la perception qu'a l'utilisateur de l'information pertinente, (2) la représentation des informations jugées et manipulées par l'utilisateur et ce, à l'aide de modèles théoriques reconnus dans le domaine, (3) la maintenance de ces profils à travers les interactions de l'utilisateur avec le système d'accès à l'information.
Ce verrou est loin d'être levé comme en témoigne les nombreuses manifestations scientifiques du domaine telles que les conférences *Information, Interaction in ConteXt (IIiX)*, *Information Retrieval In conteXt (IRIX)* qui est workshop *Context based Information retrieval* annexé à la conférence *User modelling*.

1.4.2 Décliner l'adaptation au contexte dans le processus de recherche d'information

Ce verrou est attenant au précédent en ce sens que la difficulté fondamentale posée à ce niveau est d'identifier les voies et les moyens permettant de rendre la recherche d'information plus précise en exploitant un modèle du contexte de l'utilisateur qui émet la requête. La problématique est généralement posée dans la littérature à travers la "position" utile du contexte dans la chaîne globale d'accès à l'information (Micarelli et al. 2007). Les travaux actuels dans le domaine, déclinent ce verrou à travers les questions suivantes :

- *Est ce que le contexte est utile dans la phase de formulation de la requête* ? : il permettrait alors de mieux interpréter le besoin en information, de mieux cibler l'intention véhiculée par la recherche, de préciser les préférences de l'utilisateur. Dans cette configuration, le problème fondamental est d'identifier les sources d'évidence à extraire du contexte puis les traduire sous forme d'éléments à intégrer dans la formulation de la requête.

- *Est ce que le contexte est utile dans la phase d'estimation de la pertinence de l'information en réponse à une requête posée* : le problème fondamental est d'identifier un degré de rapprochement entre l'information et le contexte défini. C'est un problème difficile pour une raison évidente liée à la différence d'univers de ces deux sources d'information.

1.4.3 Réviser les méthodologies d'évaluation de l'efficacité de la recherche personnalisée/contextuelle

L'évaluation classique des performances en RI est basée sur une approche de type laboratoire (*laboratory-based model*) initiée par *Cleverdon* (Cleverdon 1967a) dans le cadre du projet *Cranfield project II* ; c'est le cadre d'évaluation largement adopté dans les campagnes d'évaluation internationales telles

que *TREC*[5], *INEX*[6] et *CLEF*[7]. Parmi les critères de performances, le rappel et la précision sont incontestablement les plus utilisés dans les travaux du domaine ; ils permettent d'évaluer la capacité du système à présenter des documents pertinents et à rejeter des documents non pertinents. Ce cadre d'évaluation a permis l'essor de très nombreux modèles, techniques et systèmes de recherche d'information. Cependant, l'avènement de la RI contextuelle a mis en exergue les limites des protocoles d'évaluation sous-jacents et des mesures utilisées. La critique majeure était le déroulement d'expérimentations adoptant le principe d'isolation de la requête de l'utilisateur qui l'a émise. De nouvelles tâches se sont ajoutées à *TREC* depuis *2002* pour s'affranchir de cette limite mais le verrou a persisté : (1) comment intégrer le contexte de l'utilisateur dans les collections de test ? (2) comment réviser les mesures d'évaluation classiques de rappel-précision dans le sens de la prise en compte de la pertinence situationnelle (pertinence liée à la situation de recherche courante impliquant un utilisateur) et non seulement de la pertinence thématique (adéquation du contenu du document au sujet couvert par la requête) ?
C'est une large problématique qui est loin d'être résolue et sur laquelle on se penchera plus longuement dans le chapitre 5.

1.4.4 Protéger la vie privée des utilisateurs

Il est communément admis à présent que la personnalisation des services d'accès à l'information, étroitement liée à la modélisation des utilisateurs, a impact sur la vie privée de ces derniers. Les travaux précurseurs ayant pointé sur le "danger" de la personnalisation sur la vie privée des utilisateurs, sont rapportés dans *Kobsa* (Kobsa 1990) : divulgation et exploitation d'informations à des fins commerciales, de discrimination, d'espionnage etc. Les premières solutions ont essentiellement porté sur la restriction de l'accessibilité de ces modèles aux utilisateurs physiques qu'ils représentent dans un environnement local. Depuis les années *1990*, la situation a été bouleversée pour de nombreuses raisons :

- *Les systèmes d'accès personnalisés ont été portés vers le web* : les auteurs, fournisseurs de services sur le web ont pris conscience de la valeur ajoutée commerciale en rendant leurs services plus adaptés à leurs utilisateurs. Des masses de profils d'utilisateurs deviennent alors accessibles par le web, sous l'effet d'une publicité agressive, pour répondre à des objectifs commerciaux.

- *Disponibilité et diversité des sources d'information* : contrairement aux années 1980 où les principales informations exploitées pour la modélisation de l'utilisateur étaient celles explicitement déclarées par l'utilisateur, les sources sont actuellement nombreuses : historique de recherche, ses pages visitées, localisation géographique, *clicks* de souris, données psychologiques etc. La diversité de ces informations est d'autant plus importante dans les applications *ubiquitaires* où les utilisateurs sont identifiés par leur adresse *IP* et deviennent ainsi

[5]Text REtrieval Conference
[6]INitiative for the Evaluation of XML retrieval
[7]Cross Language Evaluation Forum

scrutés grâce à leur environnement physique.

– *Evolutions technologiques* : des ordinateurs plus puissants, des capteurs plus intelligents et réseaux plus flexibles sont disponibles sur le marché et permettent de collecter et traiter avec efficience des masses d'informations volumineuses.

L'ensemble de ces facteurs a ravivé le débat et la recherche autour de la protection de la vie privée. Du point de vue législatif, de plus en plus de pays mettent en place et adhérent à des chartes de protection de la vie privée qui délimitent l'accessibilité aux informations personnelles à des publics bien définis (étudiants, évaluateurs, patients, etc.), dans des situations spécifiques (d'apprentissage de modèles génériques pour des applications liées à la criminalité, aux traitements pathologiques etc.). Ces lois doivent s'accompagner toutefois de réflexions et innovations du point de vue scientifique et technologique autour de "la personnalisation et vie privée " qui est devenu un thème de d'intérêt fédérateur aussi bien dans le monde industriel que dans le monde académique comme en témoigne les *workshops* sur *privacy in context-aware systems* annexés aux conférences *UbiquitousComputing*[8] et workshops sur le thème *Privacy-Enhanced Personalization* dans la conférence *User Modeling Conference*.
L'objectif fondamental de ces recherches est de trouver le compromis entre la personnalisation en tant que service permettant de fidéliser les utilisateurs et la protection de leur vie privée relativement à une diffusion élargie de leurs informations personnelles pour des fins non explicitement approuvées.
Ce verrou fait l'objet de recherches émanant de plusieurs disciplines : systèmes d'information, interaction homme-machine, recherche d'information, hypermédia adaptatif etc.

1.4.5 Passer l'échelle

Le traitement de grands volumes d'informations est souvent désigné par l'expression «passage à l'échelle». Plus précisément, le passage à l'échelle d'une technique ou d'un algorithme désigne sa capacité à traiter des volumes considérables d'informations tout en conservant une complexité du même ordre de grandeur réelle, c'est-à-dire chronométrée, que celle induite par le traitement des volumes antérieurs moins importants. Dans le domaine général de la recherche d'information, la dimension du problème et le contexte d'utilisation ont changé dans des proportions considérables depuis quelques années. En effet, l'espace de stockage augmente continuellement puisque : (1) la production de données, sous forme électronique, continue de croître, (2) les documents électroniques contiennent de plus en plus souvent des informations multimédias (images et graphiques sont courants, mais audio et vidéo tendent à se généraliser) (3) des méta-informations sont générées et associées aux données de base afin de faciliter les accès ultérieurs, (4) les utilisateurs accèdent à des sources de plus en plus vastes et disséminées, le cas extrême étant le web.
Dans le domaine particulier de la recherche d'information contextuelle,

[8] http ://www.cs.berkeley.edu/ jfc/privacy/

au volume des sources d'informations interrogées, s'ajoute le volume des méta-données portées par les modèles des utilisateurs et le cout de leur exploitation dans le processus d'accès à l'information. Ce problème de passage à l'échelle est particulièrement posé dans le cas de modèles d'accès collaboratifs impliquant une base d'utilisateurs avec des matrices de modèles creuses (Anand 2007). De nombreuses approches visant la réduction de la dimensionnalité de l'espace de représentation des utilisateurs par des techniques de décomposition par valeur singulière (Pryor 1998) ont été proposées et appliquées particulièrement à des systèmes de recommandation ; ces approches se sont toutefois avérées efficaces dans le cas de bases de dimensions plus réduites que celles utilisées dans les applications commerciales d'aujourd'hui. Le coût de l'exploitation de ces bases de modèles est d'autant plus critique qu'elles sont sujettes à évoluer continuellement en corrélation avec l'avènement de nouvelles informations. C'est un réel verrou pour la prochaine décennie (Anand 2007).

1.5 Nos investigations de recherche

Dans la lignée des travaux de recherche allant de la vision système à la vision utilisateur/cognitive en passant par une vision large du contexte de recherche, nous nous sommes intéressés depuis le début de nos activités de recherche à l'adaptation du processus de recherche d'information pour un utilisateur donné. Ces travaux se greffent à la suite de ceux menés dans l'équipe SIG, depuis les années 80, qui ont abouti à la mise en œuvre des systèmes *Infodiab*, *Etoile* puis *Mercure*, basés respectivement sur les modèles de RI booléen, vectoriel et connexioniste. Ces systèmes ont évolué aux années 2000 vers une plate-forme, en l'occurrence *RFIEC (Recherche, Filtrage d'Information et Extraction de Connaissances)*(http ://www.irit.fr/RFIEC/), qui est un support technologique permettant la capitalisation et la mutualisation d'approches et d'expériences existantes dans le domaine de l'indexation et de la recherche d'information. La plate-forme *RFIEC* est actuellement en cours de migration vers une autre plate-forme, en l'occurrence *OSIRIM (Open Services for Indexing and Research Information in Multimedia)* (http ://www.irit.fr/OSIRIM/index.php), dont le but est de proposer un environnement homogène et ouvert (*open source*) pour la recherche sur l'indexation et la recherche d'information dans des contenus multimédias.

Il est évident que nos objectifs scientifiques se sont renouvelés par positionnement de nos travaux à l'état de l'art dans ce domaine particulier de la RI adaptative, mais notre finalité est demeurée la même : *déployer des algorithmes, modèles et techniques permettant de mettre en œuvre une recherche d'information adaptée aux besoins des utilisateurs*. Si les efforts des premières années (1995-2004) ont principalement porté sur le développement d'algorithmes et stratégies automatiques ou semi-automatiques d'adaptation, le rôle de l'utilisateur dans le processus de recherche d'information est devenu, les années suivantes (2004 à ce jour), peu à peu notre sujet de préoccupation majeur et ce en accord avec les nouveaux besoins et défis posés dans le domaine de la RI se déclinant par une mutation de l'approche d'adaptation de la vision système vers la vision utilisateur.

Plus précisément, notre projet de recherche se décline par une composition de briques élémentaires inter-connectées visant chacune d'elles la résolution, en partie, de principaux verrous posés par la RI contextuelle. Comme illustré sur la figure 1.3., nos travaux de recherche ont particulièrement tenté de lever les trois premiers verrous présentés dans la section 1.4. La première brique de nos travaux constitue une contribution à la définition des profils sources et profils cognitifs des utilisateurs (notée A), définissant ainsi le contexte de recherche considéré comme le référentiel des modèles d'accès qui vont s'y greffer. A ce titre, la seconde brique de nos travaux (notée B) exploite le contexte de recherche ainsi défini pour mettre en œuvre des techniques qui améliorent l'interprétation du besoin en information et des modèles qui permettent de mieux évaluer sa pertinence vis-à-vis d'une requête donnée. Enfin, la troisième brique (notée C) constitue des travaux transversaux qui ont pour objectif d'évaluer les modèles de recherche proposés déployant des protocoles intégrant le contexte comme dimension de l'évaluation de l'efficacité du SRI.
Dans la suite, nous faisons abstraction de l'ordre chronologique de déroulement de nos travaux en faisant le choix de les présenter comme des contributions pour la résolution de verrous bien identifiés. Ces contributions seront détaillées dans des chapitres qui leur sont dédiés.

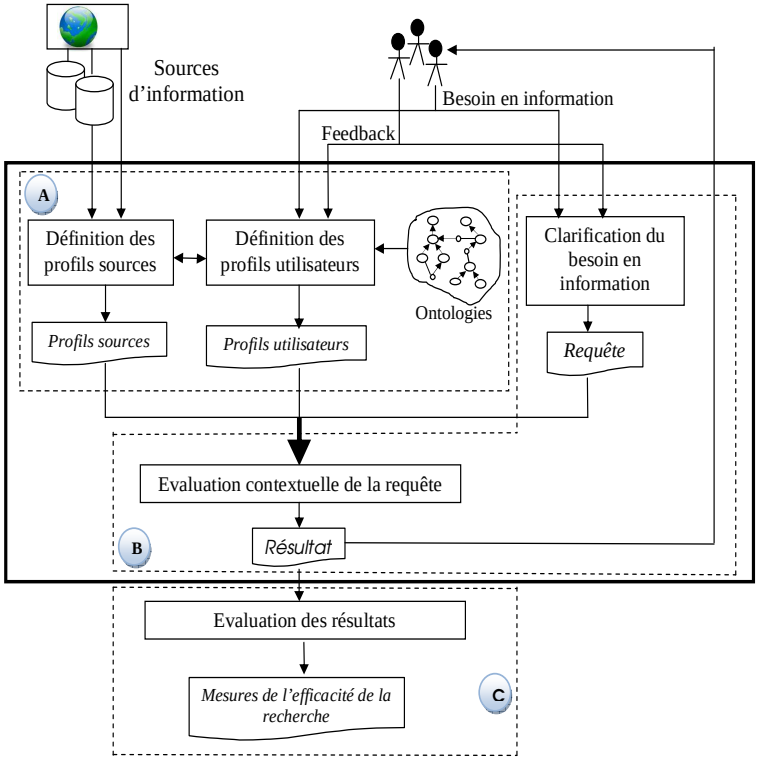


FIG. 1.3 – *Contributions à la spécification et conception d'un environnement d'accès contextuel à l'information*

1.5.1 Verrous ciblés

Même si la notion de contexte n'a explicitement germé qu'après maturation de nombreux travaux de la RI adaptative issue de l'approche orientée-système, nous organisons cette brève présentation selon la taxonomie des

dimensions du contexte présentées dans le paragraphe 1.3.1 et précisons celles qui sont particulièrement considérées pour la résolution du verrou en question. Nous présentons alors un large aperçu des travaux du domaine ainsi que les motivations qui ont suscité nos recherches.

A. Formaliser, modéliser, représenter l'utilisateur et son contexte

La modélisation du contexte de recherche est au cœur de la mise en œuvre de processus d'accès contextuel à l'information. Elle consiste à décrire les caractéristiques informationnelles des différentes dimensions du contexte considérées dans le but de les intégrer comme composantes du processus d'accès au même titre que le sont la requête et le document.

1. ***Les dimensions du contexte considérées***
   Nous nous sommes intéressés à un processus d'accès à l'information initié explicitement par la requête d'un utilisateur donné. C'est une démarche de recherche d'information active, contrairement à la recherche d'information proactive mise en œuvre dans les systèmes de recommandation ou de filtrage d'information, visant l'adaptation à un utilisateur unique et non à un groupe d'utilisateurs, comme abordé dans les travaux portant sur le filtrage collaboratif. De plus nous nous sommes particulièrement intéressés à l'exploitation :
   – *du côté de l'utilisateur* : des centres d'intérêt de l'utilisateur qu'il soient à long terme ou à court terme, dans le but d'ajuster la pertinence des résultats à son profil cognitif.

   – *du côté des sources d'information* : du profil des sources d'information dans le but de réduire virtuellement l'échelle des volumes explorés en ne retenant que celles qui sont potentiellement pertinentes pour l'utilisateur.

   – *du côté de l'environnement (mobile) de l'utilisateur* : de la localisation géographique de l'utilisateur. Cette dimension fait l'objet de travaux très récents pour lesquels nous développerons d'avantage nos perspectives de recherche.

   Sur cette base, les dimensions du contexte considérées sont : (1) le contexte cognitif et contexte personnel de l'utilisateur appelé également *profil*, (2) le contexte lié aux sources d'information.

2. ***Aperçu de la problématique et bilan des travaux***
   De très nombreux travaux dans le domaine ont abordé la question critique portant sur la modélisation du contexte cognitif de l'utilisateur (Mc Gowan 2003, Liu et Yu 2004, Gauch et al. 2003, Sieg et al. 2007, Teevan et Dumais 2005). En clair, l'hypothèse de base sur laquelle repose ces travaux est que l'interprétation du besoin en information de l'utilisateur ainsi que son jugement de pertinence dépend de ce qu'il "*sait déjà*" . Le comportement de l'utilisateur durant une activité de recherche d'information peut

alors être prédit à partir de ses interactions passées avec le système comme source d'évidence permettant de décrire les contours de ses connaissances. Deux critères caractérisent une approche de modélisation du contexte. Le premier critère porte sur le type des sources d'information utilisées à la base de la modélisation. Une revue de la littérature montre que les sources les plus utilisées sont les suivantes :

- comportement de l'utilisateur perçu à l'aide d'indicateurs d'évaluation implicite (Kelly et Fu 2007) tels que l'historique des *clicks*, les données de navigation (Teevan et Dumais 2005, Shen et al. 2005) et le mouvement des yeux (Joachims et al. 2007),

- les pages et sites favoris (Mc Gowan 2003),

- informations locales (Dumais et al. 2003b) et contextuelles telles que les sources accédées comme les journaux (Reuters, New York Times), les *Blog* sites et les sites de e-commerce (Budzik et Hammond 2000).

Le répertoire d'information collectées constitue un riche répertoire qui peut éventuellement être exploité par des techniques d'extraction de données (Mobasher 2007) ou d'apprentissage automatique (Webb et al. 2001) pour construire des modèles d'utilisateurs riches (Micarelli et al. 2007).
Le second critère porte, à juste titre, sur la représentation proprement dite de ces modèles. Les représentations proposées dans la littérature reposent sur de simples mots-clés (Lieberman 1995), des classes de mots-clés (Mc Gowan 2003), une hiérarchie de concepts (R.K. et Chan 2003) ou une partie d'une ontologie de référence (Liu et Yu 2004, Gauch et al. 2003, Sieg et al. 2007).

Concernant le contexte lié à la source d'information, à notre connaissance il existe peu de travaux ayant explicitement caractérisé les sources d'information sous l'angle des préférences des utilisateurs dans la perspective d'une contextualisation du processus de recherche d'information (Xie 2008). Les études menées ont en révélé l'impact sur le jugement de pertinence des utilisateurs mais sans en donner la portée pour une recherche d'information contextuelle.

3. ***Motivations***

- *Pourquoi s'intéresser au contexte cognitif de l'utilisateur ?*
  Même si les travaux relatifs à la formalisation du contexte cognitif de l'utilisateur sont abondants, le problème de la modélisation du profil de l'utilisateur demeure un problème ouvert pour de nombreuses raisons. Deux principales raisons nous ont particulièrement motivés dans ce cadre précis. La première concerne la difficulté à cerner et distinguer notamment les centres d'intérêt à court terme et centres d'intérêt à long terme de l'utilisateur, leur

utilisation dans un processus de contextualisation cible en effet des objectifs complémentaires mais différents. La plupart des travaux focalisent sur l'un ou l'autre, leur interconnexion étant peu voire pas abordée. La deuxième raison est la difficulté de couvrir la diversité des centres d'intérêt de l'utilisateur à la fois sur le volet thématique et sur le volet temporel caractérisant leur évolution. Dans ce sens, les travaux du domaine, particulièrement ceux exploitant des représentations sémantiques du contexte cognitif, contournent ce problème en proposant un profil sémantique avec des dominances de sens non explicitement identifiées mais utilisées en globalité dans le processus de contextualisation de l'information.

– *Pourquoi s'intéresser au contexte lié aux sources d'information ?*
Notre intérêt pour la formalisation du contexte lié aux sources d'information trouve son origine dans notre volonté à aborder la problématique de l'accès contextuel à l'information dans un environnement distribué. Même si les travaux en RI distribuée ont permis de faire aboutir de nombreuses stratégies et techniques de sélection de sources et de fusion des résultats, on constate que ces dernières demeurent indépendantes du profil de l'utilisateur qui interroge ces mêmes sources. Notre volonté est alors de transposer les acquis (même s'ils sont récents) d'un environnement centralisé vers un environnement distribué en concentrant nos efforts sur un aspect caractéristique d'un tel cadre qui est en l'occurrence l'exploitation de diverses sources d'information.

B. Décliner l'adaptation dans le processus de recherche d'information

A ce niveau, l'objectif est d'exploiter au mieux l'évidence issue de l'utilisateur afin d'adapter les résultats du processus d'évaluation d'une requête à son contexte précis. Nous avons tenté de lever ce verrou, en premier sous l'angle d'une approche orientée-système, sans définition explicite du contexte de recherche, ensuite sous l'angle d'une approche orientée-utilisateur mettant en jeu des contextes cognitifs explicitement construits. Dans ce sens, nous avons abordé le problème sous deux fronts.

– le premier front consistait à procéder à la clarification des besoins en information de l'utilisateur par la prise en compte de son *feedback*, de ses préférences ou encore de son intention de recherche,
– le second front consistait à évaluer la requête dans son contexte, en ce sens que l'estimation de la pertinence de l'information est subordonnée à son "utilité" vis-à-vis du contexte de recherche identifié.

1. ***Les dimensions du contexte considérées***

Du point de vue chronologique de nos travaux, ce verrou a été tout d'abord abordé, sans prise de conscience franche de ce qu'on a appelé plus tard le "contexte". En ce sens, inscrits dans l'optique des

travaux sur la RI adaptative ou flexible, nous avons particulièrement exploité en premier lieu :

- le *feedback* de l'utilisateur en termes de jugements de pertinence de l'information,

- les préférences de l'utilisateur en termes de résultats de recherche à travers une formalisation flexible de la requête.

Dans un second lieu, notre vision ayant évolué vers la notion de contexte, nous avons alors particulièrement, en partie, levé ce verrou par la prise en compte du contexte cognitif, en l'occurrence les centres d'intérêt de l'utilisateur, pour déployer un accès contextuel à l'information.

2. ***Aperçu de la problématique et bilan des travaux***
La problématique de l'adaptation du processus de recherche d'information aux besoins des utilisateurs a été abordée dès le début des années 1970 (voir résultats de la RI orientée système, section 1.2.). La question de recherche posée à ce niveau était la suivante : *comment rapprocher au mieux les résultats de l'évaluation de la requête de ceux particulièrement attendus par un utilisateur donné, compte tenu de son besoin mental en information ?* C'est une question qui a mené à de nombreuses pistes de recherche. Les directions les plus proches de celles que nous avons investies consistent à affronter le problème sous deux angles différents : (1) clarifier le besoin en information de l'utilisateur pour mieux le satisfaire, (2) réviser le principe d'évaluation de la pertinence de l'information en tenant compte de la perception de l'utilisateur.

Concernant le premier angle, les possibilités de clarification des requêtes sont diverses. Celles que nous avons adoptées vont dans le sens de la reformulation de la requête par réinjection de la pertinence, de la formalisation de requêtes flexibles et de la détection du besoin en information induit par la requête :

- *Reformulation de requêtes par réinjection de pertinence*
Les techniques de reformulation de requêtes par réinjection de la pertinence (*relevance fedback*) constituent une sous-classe des techniques d'expansion interactive de requête (Interactive Query Expansion (IQE)) (Rocchio 1971, Harman 1988, Magennis et Rijsbergen 1997). Leur principe est d'augmenter, par un processus itératif, les requêtes en exploitant l'évidence issue des documents jugés explicitement pertinents par l'utilisateur.

- *Formalisation de requêtes flexibles*
La flexibilité des requêtes a été introduite à deux niveaux. Le premier niveau porte sur les critères de recherche pour permettre l'expression des préférences utilisateur (Duncan et Kraft, Bordogna et al. 1991, Pasi 1999). Des poids numériques ont

d'abord été utilisés. Puis, des poids qualitatifs, plus simples et plus intuitifs, ont été formulés à partir de termes linguistiques tels que : important, très important (Bordogna et al. 1991). Le deuxième niveau concerne les opérateurs liant les critères de recherche : des opérateurs flous, intermédiaires entre le AND et le OR ont été proposés Boughanem et al. (2005), et des quantificateurs linguistiques tels que tous (*all*) et au moins k (*at least k*), ont été introduits dans le langage de requête (Bordogna et Pasi 2007) comme opérateurs d'agrégation flous qualitatifs.

– *Détection du besoin en information induit par la requête*
C'est une piste de recherches qui a été investie récemment (relativement aux pistes précédemment citées, plus précisément depuis l'avènement de la RI contextuelle). Son objectif est de détecter la tâche vue comme étant une activité réalisée pour atteindre un but. Nous nous sommes particulièrement intéressés, dans ce cadre, à une tâche de recherche d'information sur le *web*. Des études empiriques dans ce sens (Lee et al. 2005b, Jansen et al. 2007), ont montré que trois grandes classes de requêtes sont exprimées sur le *web* : requête informationnelle, requête navigationnelle et requête transactionnelle. Des attributs spécifiques de la requête tels que les balises et les attributs lexicaux sont particulièrement utilisés pour identifier précisément le but de la requête. Cette évidence est alors exploitée par un processus d'évaluation de requête spécifique afin d'ajuster la pertinence de l'information en accord avec son type. Les travaux du domaine exploitent particulièrement le contenu des documents, les liens et les textes d'ancrage (Kang et Kim 2003, Westerveld et al. 2002, Li et al. 2006).

Concernant le second angle portant sur la révision du principe d'estimation de la pertinence de l'information en tenant compte de l'utilisateur qui l'a émise, nous nous sommes particulièrement intéressés à l'évaluation des requêtes préférentielles et l'évaluation contextuelle (tenant compte des centres d'intérêt de l'utilisateur).
L'évaluation des requêtes préférentielles est basée sur des méthodes d'agrégation de la RI classique utilisant des opérateurs de conjonction (ou disjonction) pondérés. Les requêtes conjonctives (respectivement disjonctions) sont évaluées par des opérateurs conjonctifs (respectivement disjonctifs). Ces opérateurs peuvent être le minimum (respectivement le maximum) pondéré ou la moyenne pondérée. Cependant, ce type d'agrégation est trop restrictif. En particulier, dans le cas de requêtes conjonctives par exemple, la non satisfaction d'un seul critère par un document donné, implique que le document n'est pas sélectionné. Pour relaxer la conjonction, des opérateurs plus flexibles tels que la moyenne pondérée ordonnée (OWA (Yager 1988)) ou le minimum pondéré ordonné (OWA min (Dubois et Prade 1986)) ont été introduits. L'idée derrière ce type d'agrégation est de donner une faible importance aux poids les plus

faibles dans le vecteur d'évaluation, minimisant ainsi l'impact des faibles termes pour éviter de pénaliser un document contenant de faibles termes Boughanem et al. (2007).

L'évaluation contextuelle des requêtes a été essentiellement abordée selon la dimension cognitive ; elle consiste particulièrement à "augmenter" le modèle de recherche d'information par les centres d'intérêt de l'utilisateur (Lin et al. 2005a) ou alors réordonner les résultats de recherche par évaluation de la proximité des résultats produits par le processus classique aux centres d'intérêt de l'utilisateur (Liu et Yu 2004, Gauch et al. 2003, Sieg et al. 2007).

3. ***Motivations***

   – *Pourquoi s'intéresser à la clarification de la requête* ?
   Notre intérêt pour cette voie de solution a été motivé pour diverses raisons marquées par l'évolution chronologique des objectifs fixés dans le but de s'affranchir de ce verrou. La motivation première, inscrite dans la mouvance des travaux sur la RI adaptative issue de l'approche orientée système, était d'aller au delà des limites des techniques classiques de reformulation de requêtes qui reposaient sur une unique requête pour explorer l'espace de documents. Notre objectif était de favoriser une exploration de cet espace par une évaluation multi-requêtes basée sur de multiples versions de la même requête, exprimant chacune une facette du besoin en information. Note objectif par la suite a évolué, sous l'effet de la profusion des travaux sur la RI sémantique et RI flexible, vers une utilisation plus étendue de l'expression du besoin, utilisant en plus de la représentation simple de la requête basée sur les mots-clés, des critères de recherche explicités par l'utilisateur. Enfin, avec l'essor de la RI contextuelle, un bilan des travaux dans le domaine, sous l'angle de la formalisation du besoin en information, nous a permis de faire ressortir l'absence de mécanismes de personnalisation basés sur la tâche. Les travaux connexes, proposaient effectivement des stratégies d'adaptation du processus de RI tenant compte des caractéristiques de la requête, mais cette adaptation était orientée contenu et non utilisateur. En ce cens, que deux utilisateurs différents exprimant la même requête avaient en retour les mêmes résultats puisque seule la morphologie de la requête déterminait son type.

   – *Pourquoi s'intéresser à l'évaluation contextuelle de la requête ?*
   L'évaluation contextuelle de requêtes constitue le cœur des travaux de la RI contextuelle. Au delà de cela, notre motivation principale se justifie par notre volonté de faire aboutir nos propositions sur la spécification de modèles d'accès adaptatifs/contextuels à l'information. Ainsi, nos contributions au niveau de la clarification des requêtes et de la modélisation du contexte ont été naturellement suivies par des investigations autour de l'exploitation de ces évidences pour évaluer les requêtes

dans leur contexte. C'est ainsi qu'en accord avec nos contributions pour la formalisation de requêtes préférentielles d'une part, et de la définition de contextes cognitifs d'autre part, nous avons développé respectivement des modèles d'appariement requête-préférence-information et requête-contexte-information.

C. Réviser les méthodologies d'évaluation de l'efficacité de la recherche

La validation empirique d'un modèle, d'une technique ou d'un algorithme en RI est une fonction transversale, fondée sur le paradigme de l'évaluation comparative des résultats qu'ils produisent dans une situation de recherche simulée à l'aide de collections de test. Cette validation est traditionnellement menée en RI selon le protocole issu du projet Cranfield (Cleverdon 1967b) et repris dans des campagnes d'évaluation reconnus telles que TREC[9]. Notons que c'est ce même protocole que nous avons exploité pour valider nos contributions relevant de l'adaptation selon l'approche orientée-système. L'évaluation a constitué un réel verrou dans le seul cadre de la validation de nos contributions relevant de l'adaptation selon l'approche orientée-utilisateur. C'est précisément ce que nous adressons dans la suite.

1. ***Les dimensions du contexte considérées***
   Nous avons particulièrement évalué l'efficacité d'un processus de recherche d'information intégrant de manière isolée puis de manière coordonnée, deux dimensions : le contexte cognitif et la tâche. En cohérence avec nos contributions citées çi-haut, nous nous sommes intéressés aux trois principales tâches associées à une recherche d'information sur le *web* initiée par une requête informationnelle, navigationnelle ou transactionnelle.

2. ***Aperçu de la problématique et bilan des travaux***
   L'objectif fondamental de l'évaluation d'un SRI est d'estimer ses performances globales. Concernant les performances d'efficacité, qui est le critère d'évaluation de la pertinence des résultats (critère le plus considéré dans la littérature), l'évaluation classique est basée sur une approche de type "*laboratoire*" (*laboratory based evaluation*), issue de la vision système de la RI. Cette approche a été initiée par Cleverdon (Cleverdon 1967b) dans le cadre du projet Cranfield II et a été adoptée dans les campagnes d'évaluation de référence dans le domaine telles que TREC, INEX[10] et CLEF[11]. L'avènement de la RI orientée-utilisateur qui s'appuie sur une activité de recherche d'information en contexte, a recentré le débat autour de la notion de pertinence qui devient subjective, situationnelle et multivaluée (Borlund 2003b) plutôt que statique et dichotomique comme le suppose l'approche d'évaluation classique en RI. Des études critiques (Law et al. 2006, Kekalainen et Jarvelin 2004) ont en effet mis en exergue les limites de l'approche d'évaluation classique en termes

[9]Text REtrieval Conference
[10]INitiative for the Evaluation of XML retrieval
[11]Cross Language Evaluation Forum

de non prise en compte des effets de l'utilisateur dans l'évaluation de la pertinence, de la faible portée des mesures d'évaluation, et de l'inconsistance des démarches d'évaluation dans des situations de recherche d'information réelles. Une revue de la littérature que nous avons effectuée (Tamine-Lechani et al. 2009), révèle qu'il existe trois grandes approches d'évaluation orientée contexte. La première (du point de vue chronologique également) est celle basée sur des tâches *HARD* (Allan 2003a) et *Interactive* de TREC (Harman 1995b) qui ont introduit respectivement l'interaction et le contexte de recherche dans le processus d'évaluation. Ces tâches sont spécifiques en ce sens qu'elles permettent d'évaluer un processus de recherche d'information en contexte dans des conditions d'exploitation très particulières définies dans le cadre de la tâche. La seconde approche est celle basée sur des études de simulations contextuelles (*contextual simulation*) qui consiste à simuler une dimension du contexte telle que l'interaction et les centres d'intérêt, à partir de collections de test (Ryen et al. 2005, Sieg et al. 2007). La troisième approche, à la fois la plus adoptée et la plus couteuse en temps, est basée sur les études d'utilisateurs réels (*user study*) qui interagissent directement avec le SRI en cours d'évaluation à travers l'expression des requêtes, la navigation et/ou sélection d'information, le jugement de pertinence des informations résultant de l'évaluation de la requête etc. (Liu et Yu 2004, Speretta et Gauch 2005, Lee et al. 2005a, Ding et Patra 2007, Shen et al. 2005).

3. ***Motivations***
   Plus qu'une motivation, nos investigations dans ce cadre relèvent d'une nécessité. En effet, la revue de l'état de l'art dans le domaine montre clairement une absence de protocoles d'évaluation partageables par la communauté. Les méthodologies d'évaluation proposées sont dédiées à des technologies bien ciblées et sont dès lors non réutilisables. Notre souci à faire valider nos travaux selon des protocoles et normes reconnues d'une part et notre capital d'expérience acquis à travers nos différentes participations aux campagnes d'évaluation TREC d'autre part, nous ont alors motivés à mener des réflexions dans le sens de la révision du schéma standard d'évaluation TREC par intégration d'une dimension liée au contexte de recherche simulé. Les tâches TREC étant très spécifiques, nous nous sommes naturellement orientés vers une évaluation basée sur les contextes simulés mettant en œuvre deux dimensions du contexte : le contexte cognitif et la tâche.

1.5.2 Contributions

Ce paragraphe a pour objectif de donner les grandes lignes sur nos diverses contributions dans le domaine. Ces contributions seront motivées, positionnées à l'état de l'art puis détaillées dans les chapitres suivants. Nous faisons le choix de les organiser, dans ce qui suit, selon les verrous

préalablement ciblés.

A. Formaliser, modéliser, représenter l'utilisateur et son contexte

Nous nous sommes particulièrement intéressés à la modélisation des centres d'intérêt de l'utilisateur d'une part et d'autre part, à la modélisation des profils de sources d'information dans un environnement distribué.
Concernant la modélisation des centres d'intérêt de l'utilisateur, nous nous sommes orientés vers des techniques de construction qui visent l'inférence des centres d'intérêt à long terme à partir de la récurrence observée sur les sujets des requêtes décrivant des centres d'intérêt à court terme. L'évolution des profils y est basée sur des mesures de corrélation thématique entre sessions de recherche successives. Nous avons fait évoluer nos modèles de représentation depuis des structures vectorielles de termes vers des représentations sémantiques sous forme d'ensembles puis de graphes de concepts, s'appuyant sur une ontologie de référence.
Concernant la modélisation des profils sources d'information, nous nous sommes particulièrement intéressés aux facteurs descriptifs liés au contenu, à la fiabilité et à la fraîcheur de la source d'information. La description du profil n'est pas intrinsèque à la source, elle est relativisée à l'utilisateur qui l'exploite dans un environnement distribué.

Les résultats les plus importants sont les suivants :
- Un modèle de représentation, basé mots-clés, du contexte cognitif représenté par les centres d'intérêt à long terme de l'utilisateur (Tamine et al. 2006b, Tamine et Bahsoun 2006, Tamine et al. 2007c, Tamine-Lechani et al. 2008).

- Un modèle de représentation, basé graphes de concepts issus d'une ontologie de référence, du contexte cognitif représenté par les centres d'intérêt à long terme et centres d'intérêt à court terme de l'utilisateur (Daoud et al. 2007; 2008c;a).

- Un formalisme de représentation des profils source d'information dans un environnement distribué (Kechid et al. 2007; 2006)

B. Décliner l'adaptation dans le processus de recherche d'information

Dans le but de s'affranchir de ce verrou, nous avons investi deux directions : clarification du besoin en information de l'utilisateur et évaluation contextuelle des requêtes.
Sur le volet de la clarification du besoin en information, nos travaux ont investi diverses questions de recherche. La première, inspirée du principe de polyreprésentation et multi-évaluation des requêtes, a consisté à appliquer les principes de l'évolution génétique en vue de dériver la (les) requête(s) dite "optimale(s)", décrivant les différents aspects de la pertinence. Même si la question de l'efficience n'a pas été explicitement abordée,

nos résultats d'expérimentations sur des collections TREC ont montré l'efficacité de l'approche générale, de ses heuristiques spécifiques telles que l'application de la technique de nichage et la définition d'opérateurs génétiques augmentés par la connaissance du domaine.
La deuxième question de recherche abordée dans le sens de clarifier le besoin en information porte sur la formalisation des requêtes comportant des conditions qualitatives. Nous avons montré que notre démarche de formalisation utilisant les graphes CP-Nets permet à la fois de structurer intuitivement les requêtes et de générer automatiquement les poids associés aux critères de recherche.
Enfin une troisième question abordée dans ce cadre depuis peu, est la détection de la tâche induite par la requête de l'utilisateur lors d'une activité de recherche d'information sur le *web*. Nous avons montré que l'exploitation de l'évidence issue de la structure et du contexte de recherche permettent d'améliorer la précision de la classification.
Les principaux résultats issus de ces travaux sont les suivants :

- Une stratégie de reformulation de requêtes basée sur les algorithmes génétiques (Tamine et al. 2003, Boughanem et al. 2002a, Tamine et Boughanem 2006).

- Un formalisme de représentation flexible basé sur les graphes CP-Nets (Boubekeur et al. 2006, Tamine et al. 2007a).

- Une stratégie de détection de la tâche basée sur l'évidence issue de la typologie de la requête et du contexte d'interaction de l'utilisateur (Tamine et al. 2008b).

Concernant *la problématique de l'évaluation contextuelle des requêtes*, nous nous sommes investis dans deux voies. La première porte sur l'évaluation de requêtes comportant des préférences qualitatives. Dans la continuité de notre démarche de formalisation de telles requêtes, nous avons proposé des modèles d'appariement requête-préférences-document qui reposent essentiellement sur la proximité de graphes. La deuxième voie concerne la définition d'un modèle inférentiel d'accès à l'information mettant en œuvre un appariement requête-document-centres d'intérêt de l'utilisateur. Le modèle formel décrit la structure de l'information d'une part et un ensemble de propriétés traduisant des heuristiques en RI personnalisée, d'autre part.
Les résultats les plus importants sont :

- Des modèles d'appariement requête-préférence-information basée sur l'appariement de graphes CP-Nets (Boubekeur et al. 2007; 2008).

- Un modèle inférentiel pour l'appariement requête-centres d'intérêt-information (Tamine et Boughanem 2006, Tamine et al. 2006a; 2007b, Tamine-Lechani et al. 2008).

C. Réviser les méthodologies d'évaluation de l'efficacité de la recherche

Enfin la problématique de l'évaluation de modèles d'accès contextuel à l'information a été abordée selon des choix guidés par notre expérience dans le domaine. En effet, notre premier objectif dans cadre étant de valider l'efficacité d'un modèle d'accès intégrant les centres d'intérêt de l'utilisateur, nous avons ciblé un protocole d'évaluation qui vise l'augmentation du cadre d'évaluation TREC par une telle ressource. Nous avons opté pour une approche fondée sur la simulation des utilisateurs pour minimiser le coût de réalisation de nos expérimentations. Dans le but d'éviter le biais de l'évaluation, nous avons également défini une méthodologie de test par validation croisée, générant ainsi une multiplicité de profils. Nos contributions à la définition de cadres d'évaluation exploitables et réutilisables pour l'évaluation des performances d'un SRI contextuel à l'information sont les suivantes :

- Un protocole d'évaluation de l'efficacité d'un modèle d'accès contextuel, guidé par les centres d'intérêt à long terme de l'utilisateur et ce, par simulation de contextes à partir d'une collection TREC destinée à une tâche *adhoc* (Tamine-Lechani et al. 2008, Daoud et al. 2008c).

- Un protocole d'évaluation de l'efficacité d'un modèle d'accès contextuel, guidé par les centres d'intérêt à court terme et centre d'intérêt à long terme, par simulation de contextes et sessions de recherche à partir d'une collection TREC destinée à une tâche HARD (Daoud et al. 2008c;a).

- Une stratégie d'évaluation de l'efficacité d'un modèle d'accès contextuel, guidé par la tâche, par simulation des sessions de recherche (Daoud et al. 2008b).

1.6 Plan du document

La suite de ce document est structurée en 5 chapitres présentant une synthèse de nos travaux et les contributions apportées. Chaque chapitre contient une introduction, un aperçu de l'état de l'art pour la résolution du verrou posé, notre contribution ainsi qu'une conclusion et des perspectives.
Le chapitre 2 se situe dans le cadre de la modélisation du contexte de recherche. Une grande partie y est consacrée à la définition et de la formalisation du contexte cognitif. On y donnera alors un ensemble de définitions consensuelles puis développerons nos technologies. On y présente également, notre formalisme pour la représentation des profils des sources d'information dans un environnement distribué.
Les chapitres 3 et 4 sont consacrés à la spécification de modèles d'accès contextuel/adaptatifs à l'information. Le chapitre 3 met l'accent sur la partie en amont du modèle portant sur la clarification des requêtes. Le chapitre 4 met l'accent sur la partie en aval liée au modèle d'appariement

lui même induit par le modèle d'accès à l'information.
Le chapitre 5 est consacré à l'évaluation de l'efficacité de la recherche d'un processus d'accès contextuel à l'information. Nous y développons une analyse critique de la question de l'évaluation des performances d'un SRI dans un tel cadre puis donnons les spécificités et les détails de notre approche d'évaluation par prise en compte du contexte cognitif et contexte information selon une approche basée sur la simulation contextuelle.
Enfin nous dressons dans le chapitre 6 un bilan prospectif permettant de faire asseoir nos recherches futures.

Modélisation du contexte de recherche

2

2.1 Introduction

Un modèle de RI classique comprend fondamentalement, selon la vision système (Baeza-Yates et Ribeiro-Neto 1999), un modèle de représentation des documents et requêtes ainsi qu'un modèle d'appariement requête-document. La vision utilisateur introduisant une nouvelle dimension liée à l'utilisateur et son contexte, y intègre alors en plus, un modèle du contexte de recherche (Ingwersen et Jarvelin 2005), dûment représenté et/ou exploité comme ressource dans la modélisation de l'appariement requête-contexte-document. Nous avons décrit dans le chapitre 1 (section 1.3) les facteurs potentiels du contexte abordés peu ou prou en RI. Notre analyse des travaux du domaine révèle que même si les études ont montré l'impact de chacune de ces dimensions sur l'utilisateur en situation de recherche d'information, les résultats se restreignent parfois à des recommandations pour la conception de processus de recherche d'information sensibles au contexte. A notre connaissance, les travaux du domaine ayant apporté à la communauté des plus values tangibles (en termes de modèles, de stratégies, de techniques et d'algorithmes) ont essentiellement porté sur les dimensions suivantes :

- *Contexte utilisateur et tâche/problème* : ces dimensions constituent le cœur des travaux de la RI contextuelle qualifiée également de *RI personnalisée*[1] supportant des représentations implicites ou explicites de l'utilisateur ou d'un groupe d'utilisateurs.

- *Moyen d'accès à l'information et contexte spatio-temporel* : ces dimensions ont été prises en compte dans le domaine particulier de la RI contextuelle dans des environnements mobiles.

Le tableau 2.1 donne une vision synthétique des travaux du domaine catégorisés selon des groupes de facteurs associés précisément à ces dimensions du contexte.

[1] Hormis quelques nuances liées à la prise en compte du contexte applicatif (Micarelli et al. 2007), on n'identifie pas dans la revue de l'état de l'art une différence particulière de sens entre les deux expressions (Anand 2007), ce point est de nouveau abordé dans le chapitre 3

Pour notre part, considérant la taxonomie des dimensions du contexte présentées dans le chapitre 1, nos activités de recherche portant sur la modélisation du contexte portent essentiellement sur la prise en compte de deux dimensions : (1) le contexte cognitif, plus particulièrement les centres d'intérêt de l'utilisateur en qualité d'individu isolé de son groupe, (2) le contexte information, plus particulièrement les sources d'information interrogées dans une activité de recherche d'information. Comme illustré sur la figure 2.2., la plate-forme d'accès contextuel et distribué à l'information déploie les principales fonctionnalités suivantes :

1. *Définition du profil utilisateur* : c'est une étape fondamentale qui consiste à caractériser l'utilisateur lors de son activité de recherche d'information. Nous rapportons en section 2.3. une synthèse de nos travaux portant particulièrement sur la modélisation de ses centres d'intérêt (partie de son profil).

2. *Définition du profil source d'information* : traduit l'étape de caractérisation de la source d'information du point de vue du contenu et de la qualité ; nos réflexions dans ce sens sont développées en section 2.4.

3. *Recherche d'information contextuelle* : correspond à l'utilisation de l'évidence issue du profil de l'utilisateur d'une part et/ou profil de la source d'information d'autre part, en vue de fournir des résultats personnalisés. Nos travaux dans ce sens seront développés dans le chapitre 4.

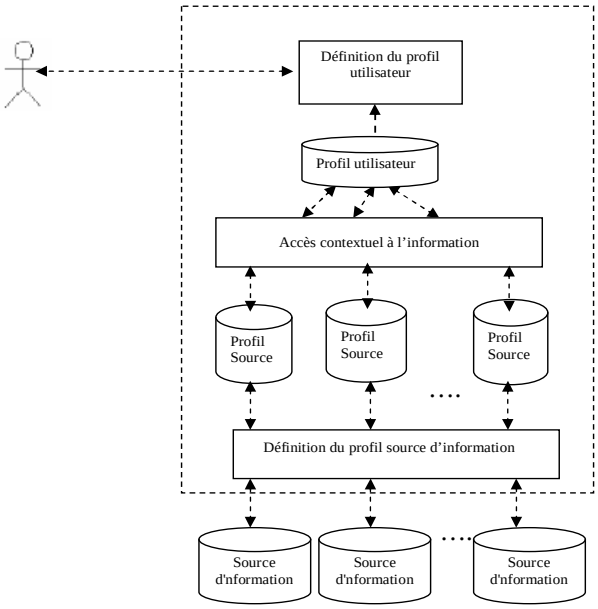


FIG. 2.1 – *Plate-forme d'un SRI contextuel dans un environnement distribué*

La section 2.2 présente une synthèse des travaux du domaine portant sur la modélisation des centres d'intérêt de l'utilisateur et des sources d'information. La section 2.3 présente notre contribution à la modélisation du contexte cognitif de l'utilisateur. Nous y décrirons notamment les évolutions de nos travaux depuis une modélisation des centres d'intérêt basée

mots-clés vers une modélisation sémantique utilisant une ontologie de référence. La section 2.4 décrit notre contribution pour la modélisation des sources d'information ; nous porterons un intérêt particulier à la prise en compte de cette modélisation dans un environnement distribué. La section 2.5 synthétise nos contributions à ce domaine de recherche. Enfin la section 2.6 conclut le chapitre et présente les perspectives de recherche.

2.2 Synthèse des travaux du domaine

2.2.1 Modélisation des centres d'intérêt de l'utilisateur

La tâche de modélisation des centres d'intérêt de l'utilisateur, qui est fondamentale dans un système d'accès contextuel/personnalisé à l'information, est plus communément appelée tâche de modélisation du profil de l'utilisateur (*user profiling*) ou simplement modélisation de l'utilisateur (*user modeling*). Dans le domaine de la RI, la modélisation du profil de l'utilisateur peut être caractérisée par trois éléments clés :

- *Durée de vie des centres d'intérêt* : selon ce critère on distingue les centres d'intérêt à court terme des centres d'intérêt à long terme. Les centres d'intérêt à court terme décrivent un intérêt de courte portée lié à la session de recherche en cours (Dumais et al. 2003b, Gauch et al. 2003, Shen et al. 2005). L'identification du centre d'intérêt requiert dès lors la définition d'heuristiques pour caractériser le basculement dans les sujets d'intérêt d'une session de recherche à l'autre (Sriram et al. 2004, Jansen et al. 2006). Les centres d'intérêt à long terme traduisent en revanche des sujets d'intérêt récurrents de l'utilisateur, des domaines d'expertise dans le cas d'application en RI dédiées (Tan et al. 2006, Liu et Yu 2004, Sieg et al. 2007). Notons que des stratégies d'évolution sont généralement sous-jacentes aux stratégies de construction des centres d'intérêt à long terme.

- *Sources d'informations* : les sources d'information sont tout d'abord déterminées par l'approche de modélisation. En effet, la source peut être simplement l'utilisateur dans le cas d'une approche de modélisation explicite (Pazzani et al. 1996). Autrement, dans le cas d'une approche implicite, la source peut être l'historique des *clicks*, les données de navigation (Teevan et Dumais 2005, Shen et al. 2005), le mouvement des yeux (Joachims et al. 2007), les pages et sites favoris (Mc Gowan 2003), les applications locales (Dumais et al. 2003b).

- *Modèle de base pour la représentation* : de très nombreux modèles ont été proposés pour la représentation des centres d'intérêt de l'utilisateur. Les plus simples sont ceux basés sur un vecteur de mots-clés (Pazzani et al. 1996), un ensemble de vecteurs mot-clés (Chen et Sycara 1998), des classes de mots-clés (Mc Gowan 2003). Des modèles plus riches sont également proposés, basés sur les réseaux sémantiques (Gentili et al. 2003), les hiérarchies de concepts (Begg et al. 1993) issus des documents consultés, ou par appariement avec une ontologie de référence (Liu et Yu 2004,

Speretta et Gauch 2005, Sieg et al. 2007).

Notre approche de modélisation du profil utilisateur a pour objectif de construire les centres-d'intérêt de l'utilisateur à partir des sessions de recherche. L'approche de construction est implicite, basée sur des sources d'évidence représentées par les contenus des documents jugés explicitement/ou implicitement pertinents par l'utilisateur. Notre contribution dans ce cadre sera étayée dans la section 2.3.

2.2.2 Modélisation du profil des sources d'information

La caractérisation d'une source d'information est une question de recherche abordée particulièrement par les communautés bases de données (BD) et RI. Les angles de vues, à ce jour, sont différents, dépendant de l'objectif de cette caractérisation.
En BD, de nombreux travaux abordent particulièrement l'enjeu de la qualité des systèmes d'information devenu important en raison de la diversité des volumes de données et des exigences de compétitivité (Naumann et al. 1999, Paiattini et al. 2006, Bouzeghoub et Peralta 2004). Dans ce sens, outre son contenu, une source d'information est caractérisée par sa qualité, définie sur le modèle support et sur les données qui les instancient. Plus précisément la qualité des données, peut être déclinée par un ensemble de facteurs tels que la fraîcheur, la fiabilité, l'exactitude et la complétude.
En RI, le facteur fondamental de qualité est la pertinence de la source pour la requête en cours d'évaluation. Les autres facteurs de qualité, sont à notre connaissance, en marge des enjeux de recherche actuels. Une source d'information est ainsi essentiellement caractérisée par son contenu dans l'objectif primaire d'y sélectionner l'information pertinente en réponse à une requête. Dans le cas particulier d'environnements distribués, le facteur de pertinence est le facteur discriminant entre les sources dans l'étape spécifique de sélection des sources devant être effectivement interrogées par le médiateur du SRI. Dans un tel environnement, la source est explicitement représentée comme un super-document basé généralement sur le modèle vectoriel comme dans les systèmes CORI (Callan et al. 1995) et STARTS (Gravano et al. 1997) et dont la qualité se mesure par son degré d'appariement avec la requête en cours. Les travaux se différencient alors essentiellement par les modèles et stratégies de prédiction de la pertinence tels que ceux basés sur la théorie de la décision (Fuhr 1999) et de l'optimisation (Si et Callan 2004).
Pour notre part, nous abordons le problème de la pertinence de l'information dans un environnement d'accès distribué à l'information en y intégrant un profil de source caractérisé par des facteurs de contenu et des facteurs de qualité et ce, dans la perspective de mettre en œuvre un accès contextuel à l'information. A notre connaissance, peu de travaux ont abordé la modélisation du profil de la source d'information sous cet angle. On rapporte particulièrement les travaux portant sur la conception de méta-moteurs tels que *Inquirus* (Glover et al. 2001) ; cependant, à défaut de profils sources explicitement définis, les moteurs de recherche qu'il déploie (*Google, Altavista, Yahoo* etc.) sont catégorisés relativement à

des profils de besoins en information prédéterminés (tels que : recherche bibliographique, recherche de sites d'accueil, recherche de nouvelles sur des évènements récents etc) servant de base à la réécriture des requêtes des utilisateurs.
Dans la section 3.4. nous résumons notre contribution pour la définition du profil d'une source d'information dans un environnement distribué.

2.3 Modélisation du profil de l'utilisateur

Selon notre approche, le profil de l'utilisateur comprend ses centres d'intérêt à court terme et centres d'intérêt à long terme. L'hypothèse qui a servi de base à nos réflexions est fondée sur l'idée d'exploiter la récurrence des sujets abordés ponctuellement par les requêtes de l'utilisateur pour construire ses centres d'intérêt général. Dans ce sens notre objectif est double :

1. identifier les différents sujets d'intérêt de l'utilisateur en se basant sur des sessions de recherche successives traitant d'un même sujet,

2. relier ces sujets pour caractériser ses centres d'intérêt généraux.

De part cette hypothèse, on se démarque de prime abord des autres travaux par : (1) la prise en compte conjointe des centres d'intérêt à long terme et centres d'intérêt à court terme (2) la modélisation puis l'exploitation de la diversité de ces centres d'intérêt.

Nos travaux dans ce cadre ont essentiellement évolué du point de vue du modèle de représentation. Nous avons en premier lieu proposé un modèle de représentation basé sur les matrices de termes puis fait évoluer notre modèle vers des graphes de concepts issus d'une ontologie de référence. Cette différenciation se décline également par les sources d'évidence utilisées pour la caractérisation des sessions de recherche. Nous décrivons dans ce qui suit l'essence de nos contributions en justifiant leur évolution.

2.3.1 Définition d'un profil bidimensionnel basé sur les matrices de termes

Le profil de l'utilisateur est multidimensionnel (Tamine-Lechani et al. 2008, Tamine et al. 2006b; 2007c), décrit plus précisément, par deux dimensions. La première représente l'historique de ses interactions avec le SRI ; elle est exploitée pour inférer la seconde dimension représentée par les divers centres d'intérêt de l'utilisateur représentés par des matrices de termes (Tamine et al. 2007c). Les deux dimensions évoluent corrélativement au cours des sessions de recherche. Dans un premier temps, nous définissons une session de recherche par l'association d'une requête et d'un ensemble de documents issus de son évaluation, jugés explicitement ou implicitement par l'utilisateur.
Plus précisément, notre procédé de définition du profil se décline en un cycle comportant deux principales étapes. La première étape consiste à représenter puis faire évoluer l'historique des interactions de l'utilisateur

avec le SRI par agrégation des informations collectées à partir de ses sessions de recherche successives.
La seconde étape a pour but de construire puis faire évoluer les centres d'intérêt de l'utilisateur en se basant sur la dimension *historique des interactions*. Plus précisément, on détermine des périodes d'apprentissage qui définissent des jalons pour l'extraction de centres d'intérêt à court terme à partir des informations agrégées dans l'historique des interactions. L'évolution des centres d'intérêt est alors basée sur une mesure de corrélation thématique qui évalue le degré de changement entre centres d'intérêt extraits durant des périodes successives.
Les paragraphes qui suivent donnent nos définitions de notions clés qui sont à la base de notre approche puis synthétisent la méthodologie de construction des centres d'intérêt.

A. Définitions

- *Définition 1*. Le coefficient de pertinence d'un terme t dans un document d à l'instant s noté $CPT^s(t,d)$ permet de décliner la pertinence relative d'un terme compte tenu des jugements de pertinence qu'il a émis et qui sont supposés êtres des indicateurs de son centre d'intérêt courant. L'expression de ce coefficient est fondée sur l'hypothèse qu'un terme est d'autant plus important pour l'utilisateur qu'il cooccure avec les termes qui lui sont *familiers* en ce sens qu'ils sont présents dans des documents déjà jugés. Les dépendances entre termes associés à des documents préalablement jugés sont vues comme des règles d'association (Lin et al. 1998).

- *Définition 2*. L'opérateur d'agrégation des sessions de recherche, noté $\oplus$, est un opérateur appliqué aux matrices représentatives des sessions de recherche dans le but de les agréger au cours du temps. A cet effet, on propose la définition d'un opérateur d'agrégation qui combine pour chaque terme son poids classique dans le document et ses poids atténués par les coefficients de pertinence calculés lors des sessions de recherche passées.

- *Définition 3*. Un centre d'intérêt à court terme (courant) traduit un besoin en information à court terme exprimé sur une courte période d'interactions avec le SRI ; il est représenté par un vecteur de termes pondérés, ordonnés par leur degré de représentativité, extrait à partir de la matrice représentative de l'historique.

- *Définition 4*. Un centre d'intérêt à long terme traduit un besoin en information récurrent observé à travers l'historique de recherche de l'utilisateur sur des sessions non nécessairement successives dans le temps.

B. Méthodologie de construction des centres d'intérêt

On se base sur l'hypothèse qu'un utilisateur a divers centres d'intérêt et qu'il peut basculer d'un centre d'intérêt vers un autre en cours de sessions

de recherche successives. Le procédé de construction est fondé sur un cycle de trois principales étapes :

1. *Historisation de l'information issue des sessions de recherche* : cette étape est réalisée en combinant l'importance intrinsèque de l'information portée par le contenu des documents jugés par l'utilisateur et l'importance relative de cette même information en tenant compte du profil courant de l'utilisateur. Pour cela, on définit formellement l'opérateur d'agrégation comme suit :
   $H^0(d,t) = S^0(d,t)$

$$H^{s+1}(d,t) = H^s(d,t) \oplus S^{s+1}(d,t) = \begin{cases} \alpha * w_{t,d} + \beta * S^{s+1}(d,t) \; si \\ t \notin T(R_u^{(s)}) \\ \alpha * H^s(d,t) + \beta * S^{s+1}(d,t) \; si \\ t \in T(R_u^{(s)}) et \; d \in R_u^{(s)} \\ H^s(d,t) \; sinon \end{cases} \quad (2.1)$$

   $(\alpha + \beta = 1)$, $s > s_0$
   La définition de l'opérateur $\oplus$ est fondée sur l'hypothèse que les termes associés aux centres d'intérêt de l'utilisateur sont récurrents. L'idée est alors d'affiner les descripteurs des documents déjà jugés par :
   – expansion éventuelle avec des termes associés présents dans des documents pertinents,
   – combinaison de l'importance classique de ses termes (relativement à la collection de documents) et de leur pertinence relative au profil, calculée à l'aide du coefficient $CPT(t,d)$ en cours des sessions de recherche passées.
2. *Détection du basculement dans le centre d'intérêt à court terme* : dans le but d'atteindre cet objectif, on applique un opérateur qui détermine la corrélation thématique entre centres d'intérêt à court terme extraits à deux périodes successives de l'historique de recherches de l'utilisateurs. Le basculement dans le sujet couvert par la requête se traduit à notre sens, dans une redistribution de l'importance relative des termes dominants associés. Du point de vue formel, on exploite des opérateurs statistiques pour la mesure des corrélations des rangs des termes dans les descripteurs des centres d'intérêt à court terme.
3. *Apprentissage des centres d'intérêt à long terme* : les centres d'intérêt à long terme sont dérivés selon le principe de récurrence des centres d'intérêt à court terme. Formellement, on définit un seuil statistique ΔI aux corrélations thématiques entre centres d'intérêt, notées ΔI et l'utilisons comme base de détection de nouveaux centres d'intérêt ou indicateurs d'évolution de centres d'intérêt déjà identifiés. Plus précisément, on procède comme suit :
   (a) $\Delta I > \sigma$. Les sessions de recherche sont inscrites dans le même contexte : pas d'indication sur l'évolution des centres d'intérêt de l'utilisateur ;
   (b) $\Delta I < \sigma$. Détection d'un changement de contexte ; deux configurations se présentent : découverte d'un nouveau centre d'intérêt

ou évolution d'un autre préalablement découvert. On procède alors de la manière suivante :

- sélectionner $c^* = argmax_{c \in I^s}(c \circ cc)$, $\circ$ est l'opérateur statistique de corrélation thématique
- si $cc \circ c^* > \sigma$ alors :
  - affiner le descripteur du centre d'intérêt c^*,
  - mettre à jour la matrice historique par élimination des lignes les moins récemment recalculées
- si $cc \circ c^* < \sigma$ alors :
  - élargir la librairie des centres d'intérêt,
  - réinitialiser la matrice historique de manière à privilégier l'apprentissage de ce nouveau centre d'intérêt, poser $s_0 = s$.

2.3.2 Vers un profil sémantique basé sur les graphes de concepts

L'approche de modélisation du profil décrite ci dessus présente les bases de définition d'un profil bidimensionnel intégrant une dimension temporelle au processus de construction des centres d'intérêt. Cependant, la diversité des centres d'intérêt étant fondée sur une mesure de corrélation de rangs des termes entre des contextes d'usages successifs, il est difficile de décliner précisément la sémantique des sujets abordés par les requêtes et par conséquent l'opportunité de la diversité détectée. Par ailleurs, la variation des centres d'intérêt de l'utilisateur, décelée à travers les requêtes qu'il a émises, ne présentent pas forcément des régularités prévisibles ; ainsi, la méthode statistique proposée serait confrontée à un risque d'erreur difficilement mesurable. Même si ce risque pourrait être amoindri en réduisant au mieux ces périodes, une piste intéressante est de mener une réflexion plus poussée sur un compromis entre les différents paramètres qui régulent l'évolution des centres d'intérêt d'un utilisateur.

Ce bilan critique nous a amenés à faire évoluer notre approche de modélisation du profil vers (Daoud et al. 2007; 2008c;a) :

1. une représentation sémantique, basée sur l'utilisation d'une ontologie de référence, permettant de traduire le sens véhiculé par le sujet de la requête et ce qui découle de sa récurrence dans le temps. Un profil est défini au niveau i de l'ontologie par un ensemble de concepts (Daoud et al. 2008c) $P_c = \left\{C_i^1, C_i^2, \ldots, C_i^n\right\}$ ou alors un ensemble de graphes de concepts (Daoud et al. 2008a) $P_g = \left\{G^1, G^2, \ldots, G^n\right\}$ où P_g est le profil construit par extension du graphe P_c par application d'un algorithme de propagation d'évidence sur les liens hiérarchiques et liens de référence de l'ontologie,
2. un principe de définition de jalons pour l'activité de recherche d'information permettant d'identifier des sessions de recherche, non pas sur la base du temps mais sur la base du sujet ciblé.

Une étape préliminaire de représentation de l'ontologie de référence détaillée dans (Daoud et al. 2008a) aboutit à une formalisation de l'ontologie comme suit :

$$O = \left\{G^1, G^2, \ldots, G^t\right\} \tag{2.2}$$

avec G^r graphe de concepts représenté par :

1. un ensemble de concepts $Co(G^r) = \{c_1^r, c_2^r, \ldots, c_n^r\}$, $c^j = \{t_1, w_{1j}), (t_2, w_{2j}), \ldots, (t_l, w_{lj})\}$ où t_i est un terme issu des documents associés au concept c^j, w_{ij} est le poids de représentativité du terme t_i dans le concept c^j,
2. deux types de relations entre concepts : (a) relation hiérarchique notée c_k *is a* c_l, (b) relation de référence notée c_k *référence* c_l, avec $c_k, c_l \in Co(G^r)$

Les étapes clés de la méthodologie de construction du profil sémantique de l'utilisateur sont les suivantes : dérivation du profil sémantique des centres d'intérêt à court terme, évolution vers des centres d'intérêt à long terme. Ces étapes sont synthétisées dans ce qui suit.

A. Dérivation du profil sémantique des centres d'intérêt à court terme

Le but de cette étape est de décrire le profil du besoin en information véhiculé par des requêtes successives traitant d'un même sujet, qualifié de centre d'intérêt à court terme de l'utilisateur. Ces requêtes sont alors inscrites, selon notre optique, dans la même session de recherche vue comme un ensemble d'activités de recherche initiées chacune par la formulation d'une requête. Pour cela, nous procédons à la représentation sémantique du profil la requête en cours d'évaluation par projection de l'information agrégée qui est issue de son évaluation sur l'ontologie puis sa désambigüisation afin de produire précisément *son profil*. La représentation formelle du profil de la requête, qui est à la base de la représentation des centres d'intérêt de l'utilisateur, est essentiellement dépendante du processus de désambigüisation. Nous en avons proposé deux :

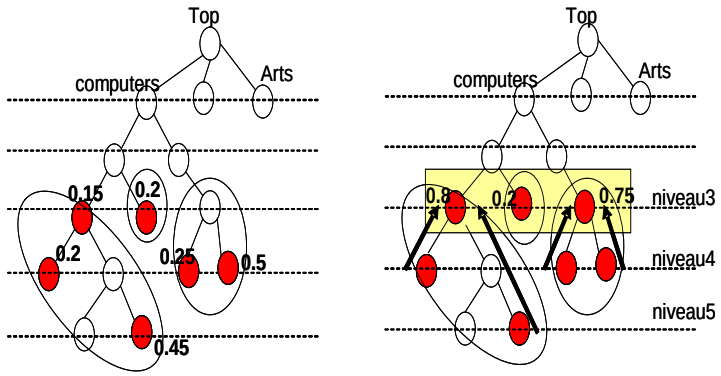
1. Désambigüisation par activation des liens hiérarchiques : le profil de la requête émis à l'instant s, soit q^s, est alors un vecteur pondéré de concepts issus de l'ontologie $\vec{q^{s+1}} = \{(c_1, sw(c_1)), (c_2, sw(c_1)) \ldots, (c_n, sw(c_n))\}$
2. Désambigüisation par activation des liens de référence : le profil de la requête est alors un graphe de concepts issus de l'ontologie, dont la représentation formelle est analogue à celle présentée en 2.2.

Ces stratégies de désambigüisation diffèrent principalement par le type de liens exploités pour le calcul du score d'activation des concepts : activation par utilisation des liens hiérarchiques et activation par utilisation des liens hiérarchiques et liens de référence.

– *Activation par utilisation des liens hiérarchiques.* On procède à une désambigüisation qui est basée sur l'hypothèse suivante : *l'importance d'un concept dans la représentativité du profil de la requête est déterminée récursivement par celle de ses sous-concepts (via la relation "est un")*. On

privilégie ainsi l'apparition de groupes de concepts liés sémantiquement plutôt que des concepts éparses dans l'ontologie. A cet effet, comme illustré sur la figure 2.1, on opère une propagation des scores jusqu'au niveau 3 de l'ontologie, en calculant pour chaque concept c_j de ce niveau ayant n sous-concepts $S(c_j)$, un score de pertinence calculé comme suit :

$$sw(c_j) = \frac{1}{n} \cdot \sum_{1 \leq k \leq n \wedge c_k \in S(c_j)} sw(c_k) \tag{2.3}$$



(a) Agrégation de concepts par niveau (b) Calcul des scores agrégés par concept

FIG. 2.2 – *Principe de désambigüisation du profil de la requête*

– *Activation par utilisation des liens hiérarchiques et liens de référence.* Nous exploitons dans cette configuration aussi bien les liens hiérarchiques "est un" que les liens non hiérarchiques correspondant aux liens de référence de l'ontologie de type "symbolique" (*symbolic*) et "relié à" (*related*). Le principe de projection du profil de requête est basé sur l'hypothèse suivante : *le profil sémantique d'une requête q est un graphe de concepts reliés sans être forcément un sous-graphe unique de l'ontologie, de manière à traduire la perception individuelle de l'utilisateur du centre d'intérêt général en cours d'identification à travers la requête q"*. Sur la base de cette hypothèse, l'algorithme de désambigüisation, présenté ci-après, parcourt les liens sémantiques puis agrège les scores de concepts en respectant la topologie du graphe de l'ontologie. A cet effet nous avons proposé de nouvelles représentations conceptuelles du profil requêtes dans le but déterminer un ordre de représentativité des concepts d'une requête donnée en corrélation avec la sémantique véhiculée par les concepts associés aux requête de la même session. En clair, on fait émerger un ensemble de concepts représentatifs du besoin en information auquel répond la session de recherche en cours.

B. ***Maintenance des centres d'intérêt.***

La maintenance porte aussi bien sur les des centres d'intérêt à court terme que sur les centres d'intérêt à long terme. La maintenance des

centres d'intérêt à court terme étant basée sur la récurrence du sujet de la requête à travers une session de recherche, nous procédons alors à la détection du basculement dans le sujet de recherche puis l'affinement du centre d'intérêt courant dans le cas où le basculement n'a pas été détecté.

Algorithme 1 Algorithme de propagation des scores des concepts

Entrée : θ^s est l'ensemble initial de concepts activés par l'information agrégée issue de l'évaluation de q
Sortie : $G_q^s = (Vs_q, Es_q)$ le graphe sémantique résultat
$\theta^s = \{c_1, c_2, .., c_n\}$, $ListGraphs = \emptyset$
Pour chaque concept $c_i \in \theta^s$ **Faire**
  $Queue_i = \{c_i\}$
  //initialisation du graphe induit par c_i
  $G_i = (V_i, E_i), V_i = V_i \cup \{c_i\}, E_i = \emptyset, w(G_i) = score(c_i)$
  Tantque $Queue_i.HasElement()$ **Faire**
    $c_j = Queue_i.PopElement()$
    //extraire les liens (*is-a*, *symbolic*, *related*)
    $\ell_j = GetLinkedConcepts(c_j)$
    Pour chaque concept $c_k \in \ell_j$ **Faire**
      Si $e_{jk} \in S$ **Alors**
        $\alpha = \alpha_S$ // arc de type *symbolic*
      Sinon Si $e_{jk} \in R$ **Alors**
        $\alpha = \alpha_R$ // arc de type *related*
      Finsi
      //propagation de scores pour tous les concepts reliés
      $score(c_k) = (\alpha * score(c_j) + score(c_k))/(\alpha + 1)$
      $V_i = V_i \cup c_k, E_i = E_i \cup e_{jk}, w(G_i) = w(G_i) + score(c_k)$
      Si $c_k \in \theta^s$ **Alors**
        $\theta^s = \theta^s - \{c_k\}$
        $Queue.PushElement(c_k)$
      Finsi
    Fin pour
  Fin tantque
  $ListGraphs = ListGraphs \cup \{G_i\}$
Fin pour
//si deux graphes induits G_m, G_n ont des concepts communs
Pour chaque $G_m, G_n \in ListGraphs$ **Faire**
  Si $V_m \cap V_n \neq \emptyset$ **Alors**
    $E_m = E_m \cup E_n, V_m = V_m \cup V_n, w(G_m) = w(G_m) + w(G_n)$ // fusionner les graphes
  Finsi
Fin pour
$G_q^s = argmax_{ListGraphs(G_i)}(w(G_i))$;

1. *Détection du basculement dans le sujet de la requête* : notre approche pour détecter ce basculement est basée sur la mesure de corrélation thématique appliquée entre le centre d'intérêt courant (dérivé à partir des requêtes précédentes), soit C^s, et le profil de la requête en cours d'évaluation, soit $\vec{q_c^{s+1}}$.

   La corrélation thématique entre le profil de la requête $\vec{q_c^{s+1}}$ et centre d'intérêt $\vec{C^s}$ est ensuite calculée puis seuillée. Une valeur inférieure à un seuil prédéterminé, signifie un basculement dans le sujet des requêtes. A l'inverse une valeur de seuil supérieure au seuil signifie que les requêtes adressent le même sujet général. La détermination de ces seuils critiques est effectuée de manière expérimentale ; ceci fait l'objet de travaux de validation résumés en chapitre 5.

2. *Affinement des centres d'intérêt à court terme.* Soient C^s et q^{s+1} les représentations sémantiques associées respectivement au centre d'intérêt et requêtes courants, la maintenance des centres d'intérêt est exprimée par des opérations de mise à jour de la représentation sémantique du centre d'intérêt courant en considérant la distribution des poids des concepts communs. Sur ce principe, on calcule le nouveau poids du concept c_j dans le centre d'intérêt à court terme C^s comme suit :

$$sw_{C^{s+1}}(c_j) = \begin{cases} \beta * sw_{q^{s+1}}(c_j) + (1-\beta) * sw_{C^s}(c_j) \; si \; c_j \in C^s \\ \beta * sw_{q^s}(c_j) \; sinon \end{cases} \tag{2.4}$$

   où $sw_{C^s}(c_j)$ est le poids du concept c_j dans le centre d'intérêt C^s, $sw_{q^s}(c_j)$ est le poids du concept c_j dans q^s.

Outre l'évolution des centres d'intérêt à court terme véhiculés par les session de recherche, comme détaillée ci-avant, l'évolution des centres d'intérêt traduit, selon notre approche l'expression de la récurrence des centres d'intérêt à court terme pour faire émerger des représentations génériques des centres d'intérêt à long terme. Des réflexions ont été menées dans ce sens et présentées précédemment dans la cadre de notre première contribution portant sur la définition d'un profil à base de matrice de termes (Cf. paragraphe 2.3.1). Nous investissons actuellement la définition de modèles des centres d'intérêt à base d'ensembles de graphes en vue de faire asseoir une représentation sémantique qui offre une flexibilité pour la mesure d'appariement requête-centres d'intérêt-document. C'est une piste de recherche que nous développons en perspective.

2.4 Modélisation du profil de la source d'information dans un environnement de RI distribuée

Une source d'information, est caractérisé par un profil lié à son contenu et un profil lié à sa qualité. Le profil lié au contenu, largement considéré par les travaux du domaine, traduit la structure d'informations contenue dans la source ; c'est un profil qui évolue, indépendamment de l'utilisateur. La spécificité de nos travaux dans ce cadre réside dans la caractérisation du profil qualité. Plus précisément, notre objectif est de caractériser différents profils pour chaque source d'information, chaque profil étant relatif aux

préférences qualitatives récurrentes de l'utilisateur quant à l'utilisation de cette source. En clair, comme tout utilisateur n'a ni la même perception, ni le même usage des sources d'information, notre objectif, est de dériver à travers l'historique des interactions utilisateur-source, un profil de source personnalisé. On décrit dans ce qui suit les éléments clés du profil source :

A. Le profil du contenu de la source

Le contenu de la source d'information S_i est défini par :

- sa structure globale donnée par le descripteur du super-document : $S_D^i = \frac{1}{N}\sum_{i=1}^{N} d_i$ où d_i est le descripteur vectoriel du document d_i obtenu selon le modèle vectoriel basé sur le schéma de pondération $tf \times idf$,
- la structure individuelle des documents qu'elle comporte donnée par un descripteur $S_F^i = (f_1, f_2, \ldots, f_k)$ défini sur les dimensions caractéristiques $(F_1, F_2, \ldots, F_k)$où F_i est une caractéristique structurelle telle que : le format, la langue, l'accessibilité etc. Les valeurs f_i quantifient le poids de la caractéristique vis-à-vis de la source considérée.

B. Le profil de la qualité de la source

Nous caractérisons la source selon deux principaux facteurs : la fiabilité et la fraîcheur.

- **Fiabilité de la source.** Nous définissons la fiabilité d'une source comme un facteur de qualité qui traduit la perception de l'utilisateur vis-à-vis de la confiance portée aux informations qu'elle comporte. Notre hypothèse de travail repose sur l'idée que la fiabilité d'une source, telle que perçue par un utilisateur donné, peut être estimée par le degré de son usage. A cet effet, elle est estimée grâce une métrique définie sur la base de l'historique de recherche de l'utilisateur, en considérant les sources d'information usagées. Pour cela, on définit l'intervalle $u = [s-1 \;\; s]$ comme le temps séparant les instants $s-1$ et s liés à deux usages successifs d'une source donnée. Ces derniers correspondent aux instants où des documents, contenus dans cette source, sont supposés être pertinents suite à l'évaluation d'une requête donnée q à l'instant s, comprenant au total DP^s documents pertinents contenus dans les sources $O = \{O_1, O_2, \ldots, O_n\}$. Pour chaque source O_i, on calcule un degré de fiabilité ponctuelle $r^s(O_i)$

$$r^s(O_i) = \frac{|< d_k, O_i >|}{|DP^s|} \tag{2.5}$$

où $d_k \in DP^s$, $|< d_k, O_i >|$ est le nombre de documents sélectionnés, provenant de la source O_i.
Le degré de fiabilité d'une source dans un intervalle d'usage u est alors calculé comme suit :

$$r^u(O_i) = \beta * r^{s-1}(O_i) + (1-\beta) * r^s(O_i) \tag{2.6}$$

$$0 < \beta < 1$$

– **Fraîcheur de la source.** La fraîcheur d'un document est définie par son âge calculé comme la période séparant sa date de création de la date (instant) d'évaluation de la requête. Dans cette optique, on relativise la fraîcheur de la source d'information par rapport au profil de la requête en cours d'évolution, en ce sens que la fraîcheur de la source est une grandeur relative et non absolue, dépendant du besoin en information courant. Plus précisément, soit une source d'information S_i, on caractérise sa fraîcheur $H_t(q)$ à la période t relativement à la requête q, en calculant la densité temporelle des n premiers documents retournés soit $D_{top}(q)$ comme suit :

$$H_t(q) = \frac{|< D^t, D_{top}(q) >|}{|S_i^t|} \tag{2.7}$$

où S_i^t est le nombre total de documents datés de la période t, $< D^t, D_{top} >$ est le nombre de documents datés de la période t parmi les documents D_{top} retournés par la requête q. Dans le but de caractériser la distribution de la fraîcheur de l'information permettant de déterminer la sensibilité de la requête à ce facteur considérant le contenu de la source S_i, on calcule une mesure de l'assymétrie de la distribution temporelle de la densité H_t comme suit :

$$Ass_i(Q) = Skew(H_t(q)) \tag{2.8}$$

où *Skew* est une fonction statistique qui est utilisée pour mesurer l'assymétrie d'une distribution de probabilités, soit : $skew(Y_1, Y_2, \ldots, Y_l) = \frac{\sum_{i=1}^{l}(Y_i - \overline{Y})}{(l-1)\times s^2}$ avec s la valeur de déviation standard, $\overline{Y}$ la moyenne des valeurs Y_i $(i = 1 \ldots l)$.
Lorsque la valeur de $Ass_i(q)$ est élevée, la source d'information S_i est sensible à la fraîcheur de l'information en accord avec la requête q. Le nombre de pics observés au niveau de chaque source devient alors un critère pour leur sélection.

2.5 Contribution au domaine de recherche

La modélisation du contexte de recherche est un problème reconnu difficile dans le processus général de conception d'un SRI contextuel (Micarelli et al. 2007, Anand 2007). Nous avons contribué en partie à sa résolution en nous fixant l'objectif d'exploiter la source d'évidence liée à l'historique de recherche de l'utilisateur dans le but de modéliser à la fois ses centres d'intérêt et une représentation des profils sources d'information particulièrement caractérisée par des facteurs qualitatifs. Dans les paragraphes qui suivent, nous mettons en évidence les spécificités de nos travaux relativement aux autres travaux du domaine puis donnons un aperçu du cadre qui a permis leur déroulement.

2.5.1 Positionnement de nos travaux vis-à-vis de la littérature

Sur le plan de la modélisation des centres d'intérêt de l'utilisateur, la synthèse présentée en section 2.2.1. met en évidence l'abondance des contributions rapportées dans la littérature particulièrement depuis les années

1990. Parmi ces travaux, soulignons que les plus proches des notres sont particulièrement ceux qui sont fondés sur une modélisation des centres d'intérêt à partir de l'historique des interactions de l'utilisateur que ce soit sur la base de représentations basées sur les termes ou sur une hiérarchie de concepts issue ou non d'une ontologie de référence (Liu et Yu 2004, Speretta et Gauch 2005, Sieg et al. 2007, Sriram et al. 2004). Du point de vue conceptuel, nous nous démarquons de ces travaux selon les principaux points suivants :

1. la distinction entre centres d'intérêt à court terme et centre d'intérêt à long terme et leur mise en relation,
2. l'identification de centres d'intérêts divers,
3. le découpage de l'historique de recherche en sessions de recherche non pas selon la dimension temporelle comme cela est proposé dans (Sriram et al. 2004) mais sur la base d'une corrélation thématique entres les sujets couverts par les sessions de recherche,
4. l'exploitation de l'ensemble des liens sémantiques fournis par l'ontologie de référence,
5. la représentation sous forme de graphes mettant en évidence la proximité des concepts caractéristiques d'un centre d'intérêt plutôt que sous forme d'un ensemble de concepts éparses de toute l'ontologie.

Du point de vue de l'impact des modèles ainsi définis sur l'efficacité de la recherche, nous avons particulièrement comparé nos modèles d'utilisateurs (présentés en 2.3.1 et 2.3.2) au modèle de référence dans le domaine, présenté dans (Speretta et Gauch 2005) après leur mise en œuvre dans le prototype *Syrix*. Les résultats issus de l'évaluation expérimentale montre nos modèles sont à l'origine de taux d'accroissement moyens de la précision de la recherche de l'ordre de 17% à 25%, les résultats sont détaillés dans le chapitre 5.

Sur le plan de la modélisation des profils source, on ne rapporte pas à notre connaissance, de travaux qui visent la caractérisation des sources d'information à travers les profils d'utilisateurs qui les interrogent. Notre travail, même s'il est encore au stade embryonnaire, peut être considéré comme un travail précurseur. Il devient nécessaire d'évaluer son impact dans le cadre particulier d'un environnement distribué.

Enfin, l'exploitation conjointe des profils utilisateurs et des profils sources dans un cadre unifié n'a pas encore fait l'objet de nos recherches. La démarche de modélisation suivie a inévitablement, du fait des ses différents composants, conduit à un développement de modèles et algorithmes présentant différents formalismes. Notre expérience dans ce domaine nous incite maintenant à proposer un cadre unificateur qui déploie l'ensemble des dimensions du contexte que nous avons considérées.

2.5.2 Structuration et support de nos travaux

Nos travaux sur la modélisation des profils des utilisateurs se sont intégrés dans un premier temps dans un cadre unifié, plus global de définition

du profil utilisateur issu du projet *APMD*[2] auquel nous avons participé. Le projet APMD (*http ://apmd.prism.uvsq.fr/*)) se situe au cœur de la problématique générale d'accès efficace à de grands volumes d'informations. Ce projet fait suite aux deux actions spécifiques (AS) CNRS, sur la personnalisation de l'information (AS98) et sur le passage à l'échelle dans les systèmes de recherche d'information (AS91) auxquelles j'ai également participé. Le but du projet est de mener une réflexion globale sur la personnalisation de l'information dans un environnement à grande échelle. Plus précisément, l'ambition y est de proposer des modèles formels capables d'intégrer des profils utilisateurs qui seront utilisés par des algorithmes robustes pour un accès et une présentation adaptatives de l'information. Nos travaux se sont poursuivis ensuite en menant des réflexions sur la représentation sémantique des profils dans le cadre des travaux de thèse de *Mariam Daoud* que je co-encadre. Ces travaux ont été publiés dans des conférences internationales dont *ACM Information Interaction in Context* (Daoud et al. 2008c) et journal international *Journal of Digital Information Management* (Tamine-Lechani et al. 2008).

Concernant la définition des profils sources d'information, nos travaux ont été réalisés d'une part dans le cadre de la thèse[3] de *Samir Kechid* et d'autre part, dans le cadre d'un projet interne à l'IRIT, en l'occurrence le BQR intitulé *Plate-Forme Multi-agent pour un accès contextuel à l'information (PADIRAC)* mené en coopération avec l'équipe *Système Multi-Agent Coopératifs (SMAC)* de l'IRIT. Outre la caractérisation des verrous posés par le processus de personnalisation de recherche dans le contexte particulier d'un environnement distribué, ce cadre de travail nous a permis d'évaluer la pertinence de l'approche par AMAS (Adaptative Multi-Agent System) pour appréhender une recherche d'information coopérative menée par des agents utilisateurs qui interrogent des sources d'informations caractérisées par une diversité de profils.

2.6 Conclusion et perspectives

Nous avons présenté dans ce chapitre une revue de nos travaux portant sur la modélisation du contexte en recherche d'information. La notion de contexte étant large comme en dénote la diversité des dimensions qu'elle couvre (Cf. chapitre 1, section 1.3), nous nous sommes particulièrement intéressés à la dimension liée aux centres d'intérêt de l'utilisateur vus comme facteurs de sa dimension cognitive et dimension liée à la source d'information. Nos travaux d'investigation dans ce cadre ont abouti à :

1. un modèle sémantique du profil utilisateur qui est une abstraction de ses centres d'intérêt dérivés implicitement à partir de ses interactions avec un SRI,
2. un cadre général d'exploitation des profils des sources d'information dans un processus de recherche d'information distribué.

[2]Projet ACI Masses de Données, *Accès Personnalisé aux Masses de Données*
[3]Thèse de Doctorat de l'Université d'Alger (Algérie)

En perspective de ces travaux, nos recherches futures investiront la problématique de la définition d'un contexte de recherche instanciable dans un environnement mobile et distribué. La mobilité induit des verrous liés à la prise en compte des coordonnées spatio-temporelles pour mesurer la pertinence de l'information vis-à-vis de la tâche en cours. La distribution des sources d'information et des services d'accès y greffe des problématiques liées à la caractérisation de ces sources vis-à-vis du service sollicité durant la tâche en cours d'évaluation. Nous abordons ces problématiques sous différents angles en vue d'une intégration dans des applications spécifiques. Plus précisément, nous nous intéressons, dans le cadre des travaux de thèse de *Ourida Bouidghaghèn* au problème particulier de la projection d'une collection de documents sur un espace de représentation spatial, en vue de délivrer une information à la fois pertinente et "locale" à la requête de l'utilisateur. La sensibilité de la requête au contexte géographique se décline par la spécificité du besoin lui même, la localisation de l'utilisateur, la disponibilité de l'information etc. Ces travaux seront intégrés en premier lieu aux problématiques de recherche posées par le projet *Quaero*[4] auquel nous participons. Ils seront ensuite transposés, en posant en plus le problème de la distribution et hétérogénéité des sources d'information, au cadre des applications médicales dans le cadre du projet *IAPA (Infrastructure d'Accès, de Partage et d'Analyses de données biomédicales* auquel nous participons. Nous y poserons particulièrement les problèmes de modélisation : (1) du contexte de recherche du praticien en termes d'expertise et de tâche (diagnostic, dissémination de l'information, collecte d'information etc.), (2) du contexte de l'information en termes de source d'information (contenu ex : dossier médical ; format : image, texte ; auteur : médecin traitant, patient), (3) de la distribution géographique de ces sources d'information. Nos contributions présentées dans ce chapitre, seront alors évaluées, voire adaptées à ce cadre d'application spécifique.

[4]Projet franco-allemand, de priorité nationale, visant le développement d'outils intégrés pour la gestion de contenus multimédias

Contexte	*Travaux de référence*	*Principaux objectifs*	*Sources d'information*	*Technologie*
Comportement de l'utilisateur	(Joachims et al. 2007, Agichtein et al. 2006, Shen et al. 2005, Teevan et Dumais 2005)	Prédiction des préférences de l'utilisateur, Inférence de l'intention de recherche	Historique des *clicks*, données de navigation, Caractéristiques de la requête, Mouvement des yeux	Apprentissage automatique, modèle de langage par analyse de la distribution de données d'interaction utilisateur-SRI
Centres d'intérêt et préférences de l'utilisateur	(Sieg et al. 2004b, Speretta et Gauch 2005, Liu et Yu 2004, Ding et Patra 2007, Jeh et Widom 2002, Lieberman 1995, Mobasher et al. 2000, Koutrika et Ioannidis 2005, Bouzeghoub et Peralta 2004)	Reformulation de requête, Réordonnancement personnalisé des résultats	Profil utilisateur, Profil général issu d'une ontologie, Documents jugés, Contexte des applications, Historique de recherche	Appariement requête-profil par des algorithmes d'apprentissage automatique, filtrage collaboratif *PageRank* personnalisé, Réinjection de la pertinence, Catégorisation des données, Modèle logique de définition de contraintes
Application	(Lee et al. 2005a, Lin et al. 2005a, Leuski 2005)	Utilisation des caractéristiques contextuelles des applications pour adapter les résultats de recherche	Application de procédures, de règles d'inférence	Modèle de tâche
Tâche	(Jansen et al. 2007, Kang et Kim 2003, Westerveld et al. 2001)	Adaptation des résultats au type des requêtes	Structure du document (Anchor, URL), Vocabulaire de la requête	Classification supervisée, semi-supervisée
Localisation, Temps, Moyen d'accès à l'information	(Chittaro 2003, Yau et al. 2003, Anderson et al. 2001, Hattori et al. 2007)	Adaptation des résultats basée sur les données du contexte mobile	Vocabulaire de la requête, lieu capté par des *sensors*, historique de recherche	reformulation de requête utilisant une ontologie de références spatiales, Appariement requête-situation spatio-temporelle

Tab. 2.1 – *Une vue synthétique des travaux de modélisation/exploitation du contexte en RI*

Clarification du besoin en information de l'utilisateur

3

3.1 Introduction

Il est communément admis, dans la communauté en RI, qu'une problématique majeure dans le domaine est la différence des univers de discours des utilisateurs et des auteurs de documents. Ceci se traduit par la différence de vocabulaire utilisé d'une part pour l'expression des contenus des documents et, d'autre part, pour l'expression des besoins en information. Les modèles classiques de sélection de l'information pertinente étant basés sur l'appariement des descripteurs des documents et des requêtes, il s'ensuit ainsi un défaut d'appariement (Crestani 2000) qui engendre une dégradation des performances de recherche. Ce constat est d'autant plus problématique quand on considère d'autres facteurs : requêtes courtes, volume d'information important, expression plus ou moins vague du besoin en information, etc. (Xu 1997). Les travaux du domaine y ont remédié selon trois principaux fronts. Le premier porte sur la réduction de la disparité du sens des termes portés dans les requêtes et documents. Les solutions proposées dans ce sens portent alors essentiellement sur des stratégies d'indexation sémantique de documents par sémantique latente (Deerwester et al. 1990), d'indexation conceptuelle (Woods 1997) utilisant éventuellement des ressources linguistiques telles que les thésaurus et ontologies (Lesk 1986, Resnik 1993). Le second front est axé sur la clarification de la requête exprimée par l'utilisateur dans le but d'identifier le besoin précis en information. Dans ce sens, les premières solutions issues de la RI adaptative selon une approche orientée-système se déclinent par des stratégies de reformulation automatique de requête, semi-automatique, interactive (Efthimiadis 1996) ou par réinjection de pertinence (Rocchio 1971, Harman 1992a) ainsi que la formalisation de requêtes flexibles (Bordogna et Pasi 1999) ; les limites de ces stratégies dans le cadre du web d'une part et l'émergence de la RI orientée-utilisateur d'autre part, ont fait émerger des solutions fondées sur l'utilisation d'autres sources d'évidence issues de l'utilisateur telles que le lieu (Göker et al. 2004), la tâche (Cheverst et al. 2000, Westerveld et al. 2002), les centres d'intérêt (Shen et al. 2005, Sieg et al. 2004b) dans le but de reformuler la requête et ainsi mieux préciser le besoin en information qu'elle induit. Enfin le troisième front porte sur l'appariement requête-document qui exploite au mieux l'implication du besoin en information à partir des documents. Les principales contributions dans ce sens, même

si elles répondent en outre à d'autres problématiques de la RI, portent sur les modèles logiques (van Rijsbergen 1986b) et modèles de langue (Ponte et Croft 1998).
Dans ce large spectre de travaux, nous focalisons, dans ce chapitre, sur nos activités de recherche traitant de la clarification du besoin en information de l'utilisateur dans le but ultime d'adapter le processus d'évaluation de la requête associée. Plus précisément, nous rapportons nos contributions dans ce domaine de recherche en apportant des solutions selon trois principales orientations.
La première orientation (du point de vue chronologique) concerne notre stratégie de polyreprésentation de la requête utilisant à la base des heuristiques d'évolution issues des algorithme génétiques (AG). Cette stratégie est apparentée à une reformulation de requêtes générant une multiplicité de représentations potentielles de la requête. Chaque représentation cible un sous-espace de l'espace de représentation de la collection de documents dans le but d'atteindre une *région de pertinence*. Le processus d'évaluation multi-requêtes ainsi induit traduit une recherche coopérative permettant d'optimiser la pertinence globale des résultats.
La seconde orientation, porte sur un autre angle ; elle concerne particulièrement la formalisation de requêtes préférentielles. On y introduit un nouveau langage de requêtes permettant d'exprimer les préférences qualitatives de l'utilisateur. La spécificité de ce langage concerne la prise en charge intuitive des préférences conditionnelles. Pour cela, nous exploitons les CP-Nets[1] pour la représentation de telles requêtes préférentielles conditionnelles.
Enfin, la troisième orientation qui relève de travaux issus des réflexions autour de la RI contextuelle, concerne l'enrichissement de la requête par la tâche ou l'intention qu'elle induit. Ceci passe au préalable par la détection de cette tâche ou intention puis la personnalisation de son processus d'évaluation. Nous nous sommes intéressés aux tâches spécifiques identifiées sur le *web*, liées aux requêtes de type informationnel, type transactionnel et type navigationnel.

Notons que la première et troisième solution peuvent très bien se superposer dans le sens où elles sont applicables à tout type de requête ambigüe et aboutirait conjointement à une requête reformulée annotée d'une classe générique de besoins que l'on exploiterait dans la phase d'évaluation proprement dite. La seconde solution, est en revanche applicable à des requêtes comportant des préférences conditionnelles qui ne sont pas particulièrement considérées par deux les autres types de solutions. Les sources d'évidence destinées à clarifier la requête étant différentes, nous développons dès lors des techniques différentes et appropriées à chacune d'elles.

L'organisation retenue pour ce chapitre est la suivante. La section 3.2 donne un aperçu de travaux connexes autour de la reformulation de requêtes, la formalisation de requêtes préférentielles et l'identification de la tâche induite par la requête. Les sections suivantes décrivent nos contributions. La section 3.3 présente notre approche de polyreprésentation de

[1] *Conditional Preference Networks*

la requête basée sur les AG. La section 3.4 détaille notre formalisme pour l'expression des requêtes comprenant des préférences qualitatives ; on y décrit notamment le principe d'utilisation des CP-Nets dans ce cadre. La section 3.5 présente notre démarche pour la détection du besoin en information induit par une requête en se basant sur deux sources d'évidence : (1) la structure morphologique de la requête, (2) le contexte de recherche de l'utilisateur. La section 3.6 met l'accent sur nos principales contributions à ce domaine de recherche. Enfin la section 3.7 conclut le chapitre et en dégage les perspectives.

3.2 Synthèse des travaux du domaine

En accord avec le contexte dans lequel s'inscrivent nos travaux, comme rapporté en introduction, nous rapportons dans ce qui suit une synthèse des travaux du domaine portant sur chacun des aspects explorés : (1) reformulation et polyreprésentation de requêtes, (2) formalisation de requêtes préférentielles, (3) détection du besoin en information induit par la requête.

3.2.1 Reformulation et polyreprésentation de la requête

La reformulation de requête consiste à réécrire la requête initiale de l'utilisateur dans le but d'adapter sa structure aux documents pertinents au besoin en information qu'elle traduit effectivement. Les techniques de reformulation de requête peuvent être classifiées en méthodes locales et méthodes globales.

Les méthodes locales ajustent une requête relativement aux documents qui sont retournés comme documents pertinents pour la requête initiale. Elles peuvent être interactives, se basant sur la technique dite de réinjection de pertinence (*relevance feedback*) décrite par un processus itératif qui consiste à enrichir la requête initiale de l'utilisateur par ajout et/ou repondération de termes sur la base de la structure des documents retrouvés par le SRI et jugés explicitement pertinents ou non pertinents par l'utilisateur. La première formalisation de cette réécriture a été donnée par Rocchio (Rocchio 1971). D'autres travaux ont, dans ce cadre, exploré les algorithmes d'ordonnancement des termes d'expansion selon leur qualité (Buckley et al. 1995, Harman 1992a, Beaulieu 1997, Boughanem et al. 1999). Les méthodes locales peuvent également être automatiques se basant sur un jugement implicite de la pertinence (*pseudo-relevance feedback*) dérivé de leur ordre dans la liste des résultats (Croft et Harper 1979). Cette stratégie pose essentiellement le problème de dérive du sujet de la requête (*query drift*) auquel des solutions ont été apportées notamment par (Mitra et al. 1998, Buckley et al. 1994).

Les méthodes globales se basent sur l'expansion de requête en s'appuyant sur des ressources linguistiques telles que les thésaurus (Qui et Frei 1993), les ontologies (Mandala et al. 1999, Baziz et al. 2005a), ou sur des techniques d'associations de termes dérivées à partir de l'analyse statistique du corpus (Schutze et J. 1997) ou des règles d'association (Haddad 2003).

Sous un autre angle, la reformulation de requête a été abordée selon une approche cognitive issue des travaux de *Ingwersen* reposant sur le principe

de poylreprésentation de l'information (Ingwersen 1994a). De manière générale, la polyreprésentation repose sur l'hypothèse que les informations (documents et requêtes) sont classifiables selon un ensemble de critères (considérés comme des dimensions de l'espace de représentation de l'information) et que la prise en compte de l'ensemble de ces critères permet de mieux subordonner la pertinence de l'information à la situation de recherche globale. Plus particulièrement, la polyreprésentation de la requête évoque une réécriture de la requête initiale selon différentes structures déclinant chacune d'elles un "aspect" des résultats attendus. Cette réflexion est à la base des travaux de (Belkin 1993, Ingwersen 1994a, Lee 1998) sur l'utilisation d'index multiples pour les requêtes, de différentes stratégies de recherche et combinaison de différents algorithmes d'injection de la pertinence.
C'est particulièrement sous cet angle que nous avons abordé en partie la clarification du besoin en information de l'utilisateur. A cet effet, nous avons exploité des stratégies évolutionnistes issues de la théorie des AG que nous avons adaptées à la spécificité de la tâche de reformulation de requête en RI. Nos activités de recherche dans ce cadre sont présentées en section 3.3.

3.2.2 Formalisation de requêtes préférentielles

Une requête préférentielle est une requête qui exprime les préférences de l'utilisateur sur les résultats de la recherche. L'expression de ces préférences est généralement effectuée à deux niveaux de description. Le premier niveau est global, il porte sur les conditions d'agrégation des termes de la requête, considérés comme des critères de recherche ; le second niveau, plus spécifique, porte sur l'intérêt accordé à chacun de ces critères.
Concernant le premier niveau, des opérateurs d'agrégation linguistiques flexibles tels que *au moins n*, *la plupart de*, *tous* etc. plus simples et plus intuitifs que les opérateurs booléens classiques *et* et *ou*, ont été définis (Bordogna et Pasi 1991) et formalisés à l'aide des opérateurs de moyenne pondérée ordonnée (OWA) (Yager 1988). D'autres opérateurs tels que le *et possible* (Bordogna et Pasi 1991) permettent en outre de distinguer entre les critères essentiels et les critères optionnels.
Selon le second niveau de préférences, l'intérêt accordé aux termes de la requête est généralement exprimé à travers la notion de poids associé à chaque critère de recherche. Grâce au procédé de pondération, l'utilisateur peut fournir une description plus précise de son besoin en information. Une requête est alors définie comme une expression booléenne dont les composants élémentaires sont des couples $< t, w >$ où t est un critère de recherche et w est le poids associé. Les poids de requête ont d'abord été formalisés comme des valeurs numériques (Bordogna et al. 1991, Salton et al. 1983), puis des poids linguistiques plus intuitifs ont été définis (Bordogna et Pasi 1991). Les poids numériques de la requête indiquent une contrainte qui doit être satisfaite par la représentation des documents de la collection indexée. La nature de la contrainte imposée par le critère de sélection pondéré dépend de la sémantique associée au poids. Dans la littérature, on distingue la sémantique d'importance, la sémantique

du seuil et la sémantique de la perfection. La sémantique d'importance (Bookstein 1980) définit les poids de la requête comme des mesures de l'importance relative de chacun de ses termes. La sémantique du seuil (Buell et Kraft 1981a) définit les poids des requêtes comme des conditions à satisfaire pour chaque terme de la requête considéré dans l'appariement document-requête. Autrement dit, le seuil indique le niveau d'acceptation du degré de signification d'un terme dans un document pour qu'il soit sélectionné. La sémantique de la perfection (Bordogna et al. 1991) consiste à considérer la requête pondérée comme une description du document idéal souhaité par l'utilisateur. En associant des poids aux termes de la requête, l'utilisateur souhaite rechercher tous les documents dont le contenu satisfait ou est plus ou moins proche du besoin idéal en information représenté par la requête pondérée.

La limitation principale des poids numériques de requête est de forcer l'utilisateur à quantifier le concept qualitatif et flou d'importance alors qu'il est plus naturel d'utiliser des quantificateurs linguistique tels que *important*, *très important*, *assez important* etc. *Bordogna* et *Pasi* (Bordogna et Pasi 1991) ont défini un modèle flou de recherche dans lequel les descripteurs linguistiques sont formalisés dans le cadre de la théorie des ensembles flous (Zadeh 1975) par des variables linguistiques. Un critère élémentaire de recherche est un couple $< t, w >$ où t est un terme et w est une valeur qualitative appartenant à l'ensemble des termes de la variable linguistique *Important*. Par exemple, l'ensemble des termes de la variable linguistique pourrait être l'ensemble défini par : $T(Important) = \{important,\ trs\ important,\ assez\ important,\ peu\ important\}$.
L'ensemble de ces travaux a effectivement œuvré dans le sens de la clarification du besoin en information de l'utilisateur du point de vue de l'expression des préférences (Bordogna et al. 1991, Bordogna et Pasi 1991). Cependant, à notre connaissance, les préférences dites conditionnelles ne sont pas spécifiquement prises en compte. Comparativement aux préférences classiques, les préférences conditionnelles permettent de définir en outre des conditions qui doivent être satisfaites par les critères de recherche. C'est précisément dans cette perspective que nous proposons un formalisme d'expression des requêtes préférentielles conditionnelles en utilisant les CP-Nets. Notre contribution, dans ce cadre, est résumée dans la section 3.4.

3.2.3 Détection du besoin en information induit par la requête

L'intention véhiculée par la requête traduit la tâche (*task*) envisagée par l'utilisateur durant une activité de recherche. Cette tâche peut être appréhendée sous différents angles : l'application (Cheverst et al. 2000), le problème à résoudre (Kelly 2004), le type d'information attendu (Kang et Kim 2003). Les travaux du domaine ont largement focalisé ces dernières années sur le type d'information attendu dans le cadre particulier de la recherche d'information sur le *web*. Dans ce sens, une taxonomie des requêtes a été établie ; les requêtes sont classifiées selon trois types : requêtes informationnelles, requêtes navigationnelles ou requêtes transactionnelles (Kang et Kim 2003, Lee et al. 2005b, Jansen et al. 2007). Une

requête informationnelle est une requête classique (*ad-hoc*) qui a pour intention de rechercher des informations pertinentes autour d'un sujet. Une requête est dite navigationnelle si elle concerne un site d'accueil comme point d'entrée à un large éventail d'informations autour d'une entité. Une requête est dite transactionnelle si elle porte sur un service. Des méthodes de classification supervisées (Kang et Kim 2003) et semi-supervisées (Baeza-Yates et al. 2006) sont proposées pour identifier cette intention à partir d'une requête vue comme un ensemble de mots clés. Les méthodes sont généralement basées sur la morphologie de la requête, plus particulièrement :

- *son vocabulaire* : présence de termes interrogatifs pour les requêtes informationnelles, de mots transactionnels tels que *download*, *buy* pour les requêtes transactionnelles,

- *la distribution de ses termes dans les documents "informationnels" ou "transactionnels"* de la collection : à chaque terme de la requête est associé un coefficient de distribution dans les hyperliens, *url*, etc.

- *sa longueur* : les requêtes navigationnelles sont à titre d'exemple courtes (de longueur généralement inférieure à 3).

Le type de besoin en information ou intention étant identifiés, l'objectif est de privilégier la sélection de documents répondant le mieux à cette intention. Dans ces sens, les travaux s'orientent vers la combinaison de scores de pertinence estimée à partir de différentes sources d'évidence : les liens, les *url*, les textes d'ancre etc. (Kang et Kim 2003, Westerveld et al. 2002, Li et al. 2006).
Dans cette même perspective d'adapter le processus d'évaluation de la requête à l'intention qu'elle véhicule, nous avons proposé une approche qui combine la morphologie de la requête et le contexte de recherche de l'utilisateur qui l'a émise. Ces travaux, très récents, seront développés en section 3.5.

3.3 Polyreprésentation de la requête par algorithmes génétiques

3.3.1 Principe des algorithmes génétiques (AG)

Les AG (Holland 1975) sont des techniques de recherche stochastique ayant été appliqués avec succès à différents problèmes réels et complexes. Leur principe de fonctionnement, illustré sur l'algorithme ci-après, est basé sur l'application itérative d'opérateurs stochastiques à une populations de solutions (ou d'individus). A chaque génération du processus d'évolution, des individus de la population sont sélectionnés (selon une politique de sélection) sur la base d'une métrique d'évaluation de leur qualité (*fitness*) puis recombinés (opérateurs tels que le croisement et la mutation) de manière à générer de nouvelles solutions selon une stratégie de remplacement. Le processus d'évaluation s'arrête lorsqu'un critère est avéré.

Algorithme 2 Structure de base d'un AG

```
Générer (P(0)) ;
t = 0 ;
Tantque Terminaison-non-avérée(P(t)) Faire
  P'(t)=Sélectionner(P(t)) ;
  P'(t)= Appliquer-op-variation(P'(t)) ;
  Evaluer P(t) ;
  P(t+1)=Remplacer ((P(t), P'(t)) ;
  t=t+1 ;
Fin tantque
```

3.3.2 Motivations et objectifs

Les travaux d'application des AG's à la recherche d'information sont peu nombreux ; les premiers travaux ont été effectués par Gordon (Gordon 1991) pour la dérivation des représentations optimales de documents. D'autres travaux sur l'optimisation des requêtes ont été présentés dans (Yang et Korfhage 1993, Chen 1995, Kraft et Sadisavan 1995) et ont pour objectif d'améliorer la précision de la recherche par application de transformations génétiques classiques sur une population de requêtes.
Pour notre part, nous exploitons les techniques et concepts de l'algorithmique génétique en vue de mettre en œuvre un processus de reformulation de requêtes, motivé par les éléments de réflexion suivants :

1. Les AG's sont dotés de la propriété de *parallélisme implicite* qui permet d'orienter la recherche simultanée à travers plusieurs régions de documents.

2. Les AG's induisent de manière inhérente à leur fonctionnement, une formalisation du besoin en information par polyreprésentation, reconnue efficace en RI (Ingwersen 1994a).

3. Les AG's sont dotés d'un processus d'optimisation d'individus par préservation des *briques élémentaires*. Ceci permet de traiter les termes pertinents par combinaison et non de manière isolée comme c'est le cas dans les mécanismes de reformulation classique en RI.

4. Les AG's peuvent être augmentés par *l'heuristique de nichage* qui permet de résoudre des problèmes d'optimisation multimodaux[2]. Cette heuristique, permet dans le cas du problème d'optimisation de requêtes, d'atteindre des documents pertinents de structures non forcément similaires.

3.3.3 Approche générale

Notre approche de reformulation génétique de requête est fondée sur l'idée suivante : la collection de documents est représentée dans un espace

[2]Un problème d'optimisation est dit multimodal s'il présente plusieurs (plus d'une) solution optimale

de dimension élevée caractérisé par la présence de "régions de pertinence" non forcément proches du point de vue de la distance vectorielle. En ce sens, la fonction de pertinence, même si elle est n'est pas formalisée, présente différents pics atteignables avec des vecteurs requêtes différents (voir fig 3.1). Ceci nous a motivé vers l'utilisation de techniques avancées de l'algorithmique génétique visant l'optimisation de problèmes multimodaux, qui est en l'occurrence la technique de spéciation basée sur la formation de niches (ou groupes représentatifs) de solutions potentielles différentes. Ainsi, le processus global de reformulation de requêtes, illustré sur la figure 3.1, vise l'évolution régulée de niches de requêtes représentant des directions de recherche dans l'espace documentaire défini par la collection de documents. La manipulation génétique de ces niches de requêtes est conditionnée par le jugement de pertinence de l'utilisateur d'une part, et résultats de la recherche, d'autre part.

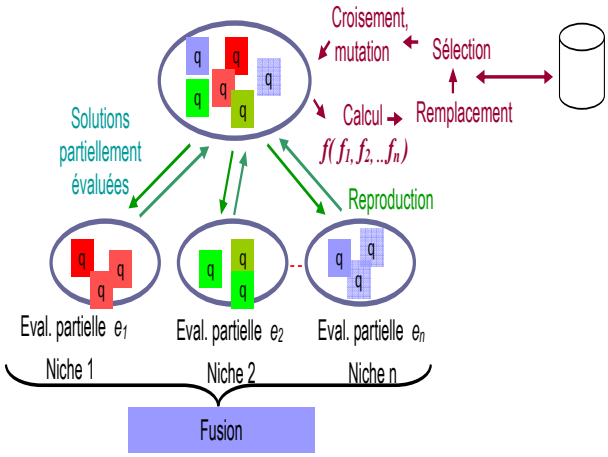


FIG. 3.1 – *Processus de reformulation de requête par AG*

3.3.4 Le modèle générique de l'AG

Le modèle d'optimisation génétique de requêtes que nous proposons (Tamine et al. 2003) est inspiré de certains comportements observés dans les niches écologiques où plusieurs algorithmes d'évaluation partielle sont déployés pour faire évoluer simultanément différentes populations (niches) de requêtes solutions. L'algorithme a pour objectif de renouveler de génération, en génération les individus requêtes en optimisant leur qualité selon un processus de sélection puis transformation basée sur le croisement et la mutation. Les différents algorithmes sont homogènes dans le sens où ils utilisent des paramètres (opérateurs de sélection et de transformation) identiques. Le principe de sélection aboutit à un échange d'individus requêtes selon un critère de décision de migration qui a pour objectif de diversifier l'espace de recherche et de retarder la convergence du processus d'optimisation. Cela permet l'obtention de solutions de meilleure qualité, mesurée par l'adéquation des résultats d'évaluation au jugement de pertinence de l'utilisateur. Pour chacune des niches, le processus de gestion des migrations intervient au terme de chaque génération de l'AG succédant à la phase de remplacement. L'ensemble de ces paramètres généraux définis sont indépendants des types de requête

et collections interrogées, ce qui rend notre modèle générique. Dans ce qui suit, nous spécifions les principaux critères spécifiques à notre modèle d'AG

A. Critère de formation des niches.

Une niche est composée d'un nombre déterminé d'individus requêtes. La composante d'une niche évolue en taille (nombre d'individus requêtes) et structure (descripteurs des individus requêtes) sous l'effet des résultats de la recherche dans la base documentaire et des échanges avec les autres niches en cours du processus d'évaluation/remplacement. La stratégie de formation des niches est basée sur la proximité de leurs résultats d'évaluation, ce qui garantit l'homogénéité des individus relativement à leur direction d'exploration de l'espace de recherche.

B. Critère de sélection des individus requêtes.

La politique de sélection des individus requêtes est appliquée à chaque niche dans l'intention de la renouveler. La stratégie aléatoire, à moindre cout, ne garantit pas la sélection des meilleures requêtes. La stratégie retenue est celle basée sur l'espérance mathématique qui offre de l'avantage de réduire l'erreur stochastique qui induit un écart très variable entre copies attendues et nombre de copies effectivement générées pour chaque individu requête. En pratique, la sélection est basée sur un mesure de score de qualité issu des mesures classiques de rappel-précision adaptée à la distribution des valeurs de scores des individus d'une même niche.

C. Critères de transformation des individus requêtes

Les transformations des individus requête appartenant à chaque niche sont caractérisées par l'utilisation d'heuristiques de reformulation de requête approuvées dans le domaine de la RI. On a particulièrement proposé des opérateurs de croisement sans point, qui visent l'ajustement des poids des termes de la requête en accord avec leur distribution dans les documents pertinents et documents non pertinents et opérateurs de mutations basés sur des fonctions sélectives des termes selon leur distribution dans la collection.

Tab. 3.1 – *Formalisation du modèle*

Concept/opérateur	***Formalisation***
Individu	$Q_u^s = (q_{u1}, q_{u2}, \ldots, q_{uT})$
Niche	$[Q_u^s \equiv_N Q_v^s] \Leftrightarrow [D(Q_u^s, Y) \cap D(Q_v^s, Y)] > Seuil$
Sélection	$P_{select}(Q_u^s) = \frac{Fitness(Q_u^s)}{Fitness_R(N_i^s)}$
Fitness	$Fitness^{(a)}(Q_u^s) = \frac{Fitness(Q_u^s)}{\|Niche(Q_u^s)\|}$
Fusion	$Rel(d_j) = \sum_{N_i \in Pop^s} \sum_{Q_u^s \in N_i^s} Fitness(Qu^s) * RSV(Q_u^s, d_j)$

D. Critère de fusion des résultats de l'évaluation des niches

La fusion des documents restitués par les individus requêtes est effectuée en deux étapes. La première étape détermine un premier ordre relatif à la supposition de pertinence, basée sur les résultats de sélection, au niveau de chaque niche. La seconde étape, réordonne alors ces listes partielles sur la base de la qualité des individus requêtes et/ou de la valeur de pertinence des documents issus de leur évaluation.

Une instance du modèle formel généré à partit du modèle générique ainsi défini est présenté dans le tableau 3.1.

Q_u^s, Q_v^s sont deux individus requêtes de la population à l'étape s de l'AG, $D(Q_u^s, Y)$ représentent les Y premiers documents restitués par la requête Q_u^s, Seuil est le nombre minimal de documents communs parmi les X premiers documents sélectionnés par chacune des deux requêtes ; $RSV(Q_u^s, d_j)$ est le score de pertinence (Relevance Status Value) du document d_j relativement à la requête Q_u^s ; $Rel_n(d_j, N_i^s)$ est la valeur de pertinence locale du document d_j dans la niche N_i^s, $Fitness\_moyen(N_i^s)$ est la valeur d'adaptation moyenne dans la niche N_i^s.

3.4 Formalisation des requêtes préférentielles basée CP-Nets

L'introduction des poids dans les termes de la requête a permis d'exprimer les préférences utilisateur sur les critères de recherche. Cependant, les approches classiques de pondération des requêtes ne permettent pas, rappelons le, de prendre en compte les préférences conditionnelles. Nous illustrons le problème posé pour les préférences conditionnelles à travers la requête qui suit : Etant donné le besoin utilisateur exprimé à travers l'énoncé suivant : "*I am looking for housing in Paris or Lyon of studios or university room type. Knowing that I prefer to be in Paris rather than to be in Lyon, if I should go to Paris, I will prefer being into residence hall (RH), whereas if I should go to Lyon, a studio is more preferable to me than a room in residence hall. Moreover the Center town of Paris is more preferable to me than its suburbs ; whereas if I must go to Lyon, I will rather prefer to reside in suburbs that in the center*".

Une telle requête fait ressortir des préférences conditionnelles. En traduisant les préférences qui y sont exprimées en valeurs numériques, une requête correspondante possible serait :

$(Paris\ \ 0.9 \wedge (RH\ \ 0.6 \vee Studio\ \ 0.3) \wedge (Center\ \ 0.5 \vee Suburbs\ \ 0.4)) \vee (Lyon\ 0.8 \wedge (RH\ 0.5 \vee Studio\ 0.8) \wedge (Center\ 0.7 \vee Suburbs\ 0.8))$.

Dans cette représentation, les poids des termes *R.H* et *Studio*, *Cente*r et *Suburbs*, sont différents lorsqu'ils sont associés avec *Paris* ou *Lyon* respectivement. Ceci traduit exactement les préférences conditionnelles exprimées dans l'énoncé du besoin utilisateur. La forme normale disjonctive de cette requête est donnée par :
$(Paris\ 0.9 \wedge RH\ 0.6 \wedge Center\ 0.5) \vee (Paris\ 0.9 \wedge Studio\ 0.3 \wedge Center\ 0.5) \vee$
$(Paris\ 0.9 \wedge RH\ 0.6 \wedge Suburbs\ 0.4) \vee (Paris\ 0.9 \wedge Studio\ 0.3 \wedge Suburbs\ 0.4) \vee$
$(Lyon\ 0.8 \wedge RH\ 0.5 \wedge Center\ 0.7) \vee (Lyon\ 0.8 \wedge Studio\ 0.8 \wedge Center\ 0.7) \vee$
$(Lyon\ 0.8 \wedge RH\ 0.5 \wedge Suburbs\ 0.8) \vee (Lyon\ 0.8 \wedge Studio\ 0.8 \wedge Suburbs\ 0.8)$.

Même si cette représentation supporte naturellement les préférences conditionnelles, elle reste problématique si quelques précautions ne sont pas prises au préalable. En effet, en supposant que chaque sous requête conjonctive de la requête globale possède un poids d'importance total, calculé par agrégation des poids individuels de ses propres termes (en utilisant l'opérateur min ou l'opérateur OWA (Yager 1988) ou simplement en moyennant les poids par exemple), on obtient un poids d'importance de *(Paris, Studio, Center)* égal à *0.56* tandis que le poids d'importance de *(Lyon, Studio, Center)* est de *0.76* impliquant que la dernière alternative est préférée à la première. Ce résultat est contradictoire avec les préférences formulées par l'utilisateur. La pondération que nous avons donnée, de façon tout à fait aléatoire et intuitive, est incorrecte. Cet exemple fait ressortir l'impact d'une pondération aléatoire ou intuitive d'une requête qualitative, sur la précision et l'exactitude de la sémantique qu'elle tente d'exprimer. Ceci illustre la tâche difficile de la pondération des requêtes qualitatives. Pour notre part, nous nous proposons de formaliser la représentation de telles requêtes à l'aide des CP-Nets, comme décrit ci après.

3.4.1 Motivations et objectifs

Un CP-Net (Boutilier et al. 1999) est modèle graphique qui exploite l'indépendance préférentielle conditionnelle dans la structuration des préférences de l'utilisateur sous l'hypothèse *ceteris-paribus*[3]. Formellement, un CP-Net est un graphe orienté acyclique, ou DAG , $G = (V, E)$, où V est un ensemble de nœuds $\{X_1, X_2, X_3, \ldots, X_n\}$ qui définissent les variables de préférence et E un ensemble d'arcs orientés entre les nœuds, traduisant des relations de dépendances préférentielles entre ces nœuds. Toute variable X_i du graphe est instanciable dans un domaine de valeurs $Dom(X_i) = \{x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ik}\}$. Le prédécesseur d'un nœud X dans le graphe est dit son parent $(Pa(X))$. L'ensemble $(X, Pa(X))$ constitue une famille du CP-Net. A chaque variable X du CP-Net, on associe une table de préférences conditionnelles $(CPT(X))$ spécifiant un ordre de préférence total sur les valeurs x_i de X étant donné chaque instance de ses parents. Pour un nœud racine, la table *CPT* spécifie un ordre de préférence inconditionnel sur les valeurs du nœud. Un CP-Net induit un graphe complet de préférences ordonné, construit sur l'ensemble de ses alternatives. Une alternative du CP-Net est un élément du produit cartésien des domaines de valeurs de ses différents nœuds. Elle est interprétée comme une conjonction de ses éléments. Un CP-Net peut être

[3]Toutes choses égales par ailleurs (all else being equal)

quantifié par des valeurs numériques de préférence, dites valeurs d'utilité, conduisant à un UCP-Net[4] dont la validité est basée sur la propriété de dominance (Boutilier et al. 2001). Les facteurs d'utilité dans un UCP-Net sont indépendants généralisés additifs (GAI) (Bacchus et Grove 1995).
Les CP-Nets ont été introduits comme outil de représentation compacte des relations de préférences conditionnelles qualitatives. A ce titre, nous les exploitons pour la représentation de requêtes préférentielles conditionnelles. De manière inhérente à leur formalisme, les CP-Nets nous permettent :

1. d'offrir à l'utilisateur un langage de requêtes graphique permettant de traduire ses préférences qualitatives,
2. de générer formellement les poids numériques correspondants conduisant ainsi à une pondération automatique des termes de la requête.

En pratique, les objectifs de nos travaux dans ce cadre, sont précisément de :

1. construire les requêtes comme des graphes CP-Nets,
2. générer automatiquement les UCP-Nets correspondants, valides du point de vue de la propriété de dominance.

Nous synthétisons dans ce qui suit l'approche que nous avons proposée dans le but d'atteindre ces objectifs.

3.4.2 Approche générale de formalisation

La formalisation repose tout d'abord sur la spécification préalable d'un ensemble de caractéristiques (variables) sur lesquelles vont porter ses préférences. Chaque caractéristique est définie sur son propre domaine de valeurs (une valeur est un terme de la requête). Pour chaque variable donnée X, l'utilisateur doit spécifier toutes ses dépendances préférentielles, ainsi que l'ordre de préférences correspondant sur $Dom(X)$. Cette description est utilisée pour construire le CP-Net requête : les nœuds du CP-Net sont les variables sur lesquelles portent les préférences utilisateur, les liens entre les nœuds définissent les dépendances préférentielles spécifiées par l'utilisateur (On supposera dans ce qui suit que le graphe résultant est un DAG). L'ordre de préférences sur un domaine de valeurs est traduit en table CPT.
La figure 3.2 illustre le CP-Net correspondant à la requête énoncée en 3.4. Les variables concernées sont *City, Housing et Place* telles que : $Dom(City) = Paris, Lyon$, $Dom(Housing) = \{RH, Studio\}$, $Dom(Place) = \{Center, Suburbs\}$. En outre, $CPT(City)$ spécifie que *Paris* est inconditionnellement préférable à *Lyon* ($Paris \succ Lyon$), tandis que $CPT(Housing)$ par exemple, spécifie un ordre de préférences sur les valeurs de *Housing*, sous la condition des valeurs prises par la variable *City* (ainsi par exemple, si *Paris* alors $RH \succ Studio$). La requête CP-Net est ensuite pondérée par des facteurs d'utilité comme décrit dans ce qui suit.

[4]Utility CP-Net

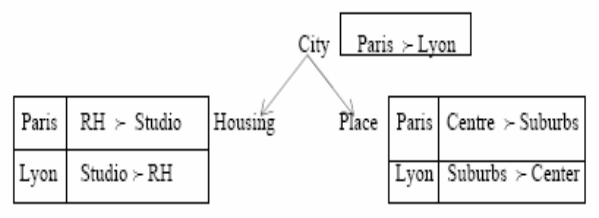


FIG. 3.2 – *Représentation CP-Net d'une requête booléenne*

Notre approche de génération automatique du UCP-Net requête est basée sur les propriétés suivantes :

1. Toute variable X doit vérifier la propriété de dominance.

2. Un ordre de préférences sur $Dom(X)$, étant donnée une valeur $u \in Dom(Pa(X))$, est traduit par une distribution uniforme des valeurs d'utilités (ou degrés de préférence) sur $Dom(X)$ étant donnée u. Intuitivement, il s'agit de distribuer uniformément des degrés de préférences sur les valeurs x_i de X, de sorte à traduire numériquement les ordres de préférence qualitatifs introduits sur les x_i dans le CP-Net considéré. Ainsi, si par exemple, une variable X, apparaît dans le CP-Net avec deux valeurs x_1 et x_2 telles que $x_1 > x_2$, ceci se traduit dans notre approche par : $f_X(x_2) = 0$ et $f_X(x_2) = 1$. Pour une variable X à trois valeurs telle que $x_1 \succ x_2 \succ x_3$, on obtient : $f_X(x_3) = 0$, $f_X(x_2) = 1/2$ et $f_X(x_1) = 2/2$. Pour respecter la propriété de dominance à la base de tout UCP-Net, on impose en outre une condition supplémentaire sur les degrés de préférences associés aux variables représentant les nœuds internes du CP-Net.

L'approche est formellement définie dans (Boubekeur 2008). Elle est fondée sur le principe de génération des valeurs de CPT en respectant les propriétés de dominance énoncées plus haut. comme suit : Soit X un nœud de la requête CP-Net, tel que $|Dom(X)| = k$, et soit $u(i)$ le degré de préférence d'ordre i (en supposant un degré de préférence croissant lorsque i croît) sur les valeurs de X.

– Pour tout nœud feuille X, nous générons les utilités sur $Dom(X)$, suivant la propriété 1, comme suit :

$$u(i) = \begin{cases} 0 \; si \; i = 1 \\ u(i-1) + \frac{1}{k-1} \;\; \forall \; 1 < i \leq k \end{cases} \tag{3.1}$$

– Tout nœud interne X, possède des descendants, et doit donc respecter la propriété de dominance (propriété 2 énoncée plus haut). Pour tout nœud interne X du CP-Net, on calcule alors la quantité : $S = \sum_i MaxSpan(B_i)$ où les B_i sont les descendants de X. Comme X doit dominer ses descendants on impose que : $Minspan(X) \geq S$.

La requête CP-Net issue de ce procédé de formalisation est évaluée à l'aide d'un modèle d'appariement approprié pour produire des résultats qui considèrent les préférences de l'utilisateur. Ce modèle d'appariement est présenté dans le chapitre 4.

3.5 Identification de la tâche induite par la requête : vers une approche basée sur le contexte de recherche

L'objectif de nos travaux dans ce cadre est de clarifier le besoin en information par la détection de l'intention "cachée" derrière la requête. Nos premières investigations dans ce sens ont porté précisément sur trois principales catégories d'intentions attachées respectivement à trois classes de besoins : informationnel, navigationnel et transactionnel. Nous décrivons dans ce qui suit l'approche générale adoptée puis développons la méthodologie formelle qui permet de prédire la classe d'une requête.

3.5.1 Approche générale pour la détection du besoin en information

Notre approche de classification du type de besoin induit par la requête repose sur l'hypothèse qu'un besoin en information est globalement véhiculé par des sessions de recherche (Tamine et al. 2008b). Chaque session comprend une requête avec un profil donné (structure, sujet, résultats) qu'il convient d'exploiter pour personnaliser les résultats. Le besoin est ainsi induit, selon notre point de vue, par l'ensemble des requêtes de la session de l'utilisateur que nous qualifions de *profil de requête* et non seulement la requête en cours. Ce point de vue est également étayé dans (Nettleton et al. 2006) où il est montré que les sessions de recherche peuvent être classifiées selon leur besoin. En outre, on suppose, que la classification d'une requête à un type donné ne peut être catégorique (ou systématique) mais probable selon ses caractéristiques morphologiques ainsi que la session à laquelle elle appartient. Pour cela, nous proposons de calculer pour chaque requête une probabilité d'être classée selon chacun des types informationnel, navigationnel ou transactionnel en considérant les deux sources d'évidence liées à la morphologie d'une part et le profil d'autre part.

3.5.2 Principe général de classification de la requête

La classification du besoin est approché comme un problème décisionnel où l'on doit maximiser la croyance qu'on a sur la classe supposée de la requête en cours d'évaluation. Partant de là, notre méthodologie générale repose sur les trois principales étapes : (1) caractérisation de la requête à l'aide de critères morphologiques spécifiques liés à chaque type de requête (2) caractérisation du profil de la requête à partir de la session de recherche courante (3) identification du besoin induit par la requête par estimation de la probabilité d'appartenance de la requête à chacune des classes en considérant sa morphologie et son profil. Formellement, cela revient à évaluer la règle de Bayes : $p(q/qf, qp)$ où q, qf et qp sont des variables aléatoires prenant des valeurs dans $\{I, N, T\}$, où I dénote une requête informationnelle, N dénote une requête navigationnelle T dénote une requête transactionnelle.

A. Description morphologique de la requête

On définit le vecteur caractéristique d'une requête Q comme :

$$F(Q) = (\|Q\|, Vb, Tr, Ti, \tau_a, \tau_t)$$

où $\|Q\|$ est la longueur de la requête (nombre de termes), Vb, Tr et Ti des valeurs booléennes indiquant si la requête contient respectivement des verbes, termes transactionnels (ex "*downloand*", "*buy*", "*purchase*" etc.) τ_a (resp. τ_t) est la proportion de termes de la requête contenus dans les textes d'ancre (resp. balise titre) dans les documents de la collection ; ces proportions, permettant de distinguer les requêtes transactionnelles des requêtes navigationnelles, sont calculées comme suit :

$$\tau_a(Q) = \frac{\sum_{t_i \in Q} \frac{nA_i}{n_i}}{\|Q\|}, \tau_t(Q) = \frac{\sum_{t_i \in Q} \frac{nT_i}{n_i}}{\|Q\|}$$

nA_i (resp. nT_i) est le nombre de documents contenant le terme t_i dans un texte d'ancre (resp. dans une balise titre)), n_i est le nombre de documents de la collection contenant le terme t_i.

Pour chaque requête, on construit un descripteur de type navigationnel $QF_N(Q)$ et un descripteur de type transactionnel $QF_T(Q)$ comme des vecteurs booléens (Tamine et al. 2008b).

B. Définition du profil de la requête

Soit Q la requête courante, n définit le profil requête $QP^{(c_i)}$ comme l'ensemble des requêtes précédentes appartenant à la même catégorie C_i. $QP^{(c_i)} = \{Q_1, \ldots, Q_{m-1}, Q_m\}$.

C. Identification du besoin induit par la requête

Cette tâche est appréhendée comme un problème de classification utilisant des sources d'évidence. Plus précisément, on définit un classifieur bayésien de requêtes dont les principales estimations sont les suivantes : soit c_i (resp. c_j) des variables aléatoires correspondant respectivement à la catégorie C_i (resp. C_j) prenant des valeurs dans $\{I, N, T\}$. On calcule la probabilité que la requête appartienne à la catégorie C_i sous condition de sa classification à la catégorie C_j en considérant ses caractéristiques comme suit :

$$p(q = c_i / qf = c_j) = \begin{cases} \alpha_i * p(qf = c_j) \ \ if \ c_i, c_j \in \{N, T\} \\ 1 - \sum_{c_j \in \{N,T\}} p(qf = c_j) \ sinon \end{cases} \tag{3.2}$$

où

$$p(qf = c_j) = \frac{\sum_{i=1..6} CF_i^{(c_j)}}{\left\| QF_{c_j}(Q) \right\|}$$

α_i est un facteur d'atténuation dépendant de la précision de la classification dans la catégorie C_i.
Cette formule suggère que plus grand est le nombre de caractéristiques

de structure satisfaites, plus importante est la probabilité que la requête Q soit de la catégorie C_j. Nous considérons que la catégorie requête informationnelle est la catégorie par défaut.

On définit la probabilité que la requête Q appartienne à la catégorie C_i sous condition que le profil requête appartienne à la catégorie C_j comme suit :

$p(q = c_i/qp = c_j) =$

$$\begin{cases} 1 - \frac{\left\|QP^{(c_i)}\right\|}{w_i} \; si \; \left[(c_i = c_j) \wedge (QP^{(c_i)} < w_i)\right], \\ \frac{\left\|QP^{(c_i)}\right\|}{2*w_i} \; si \; \left[(c_i \neq c_j) \wedge (QP^{(c_i)} < w_i)\right], \\ 1/3 \; sinon \end{cases} \tag{3.3}$$

où $\left\|QP^{(c_i)}\right\|$ est le profil requête comprenant des requêtes de la catégorie C_i, w_i est la longueur moyenne des profils requête de type C_i estimé à partir des sessions de recherche précédentes.
La formule 3.21 se base sur l'hypothèse que plus le nombre de requêtes de $\left\|QP^{(c_i)}\right\|$ de la même catégorie C_i atteint la longueur moyenne w_i, plus importante est la probabilité que la requête soit animée d'une autre intention. La probabilité de basculement à une autre intention $\left\|QP^{(c_i)}\right\|/w_i$ est uniformément distribuée sur les deux possibilités restantes. En outre, nous considérons que les intentions sont *a priori* équiprobables (1/3) lorsque $\left\|QP^{(c_i)}\right\|$ est égale à w_i.

Enfin, en se basant sur l'hypothèse d'indépendance des profils requêtes et de leur structure, on calcule :

$$p(q/qf, qp) = p(q/qf) * p(q/qp) \tag{3.4}$$

La maximisation des probabilités calculer permet d'identifier précisément la classe à associer à la requête.

3.6 Contribution au domaine de recherche

Dans le but d'aborder la problématique de la clarification du besoin en information de l'utilisateur, nous avons œuvré dans trois directions pouvant être complémentaires : (1) polyreprésentation de la requête par AG, (2) formalisation de requêtes préférentielles et (3) détection du besoin induit par la requête. Nous étayons dans ce qui suit les principaux résultats issus de chacune de ces directions ainsi que le cadre dans lequel se sont déroulées nos activités de recherche.

3.6.1 Positionnement de nos travaux vis-à-vis de la littérature

- *Polyreprésentation de requêtes par AG* : nos contributions constituent des travaux précurseurs dans le domaine, comparativement aux autres travaux du domaine (Yang et Korfhage 1993, Chen 1995, Kraft et Sadisavan 1995). Notre modèle d'AG est en effet caractérisé

par les deux principales spécificités suivantes (Tamine et al. 2003, Boughanem et al. 2002a) :

1. l'application de la technique de spéciation pour préserver la diversité des requêtes et atteindre ainsi différents pics de la fonction de pertinence. Les critères sous-jacents de renouvellement des niches et échange d'individus entre niches sont particulièrement nouveaux dans le domaine,
2. la définition d'opérateurs génétiques, non classiques, augmentés par une connaissance approuvée liée aux techniques de reformulation de requête.

Du point de vue de l'impact de notre modèle d'AG sur l'efficacité de la recherche, nous avons mené une série d'évaluations comparatives sur la collection *TREC AP88* et avons montré que :

1. l'utilisation de la technique de nichage a permis d'améliorer les performances en termes de nombre cumulé de documents pertinents, dès la 3ème génération de l'AG et ce, de l'ordre de 4*O*% relativement à une exploration classique, telle que proposée dans les autres modèles,
2. l'application d'opérateurs génériques augmentés par la connaissance du domaine permet d'améliorer les performances de l'ordre de 15% comparativement à des opérateurs aveugles.

– *Formalisation de requêtes préférentielles basées sur les CP-Nets* : la formalisation de requêtes préférentielles a fait l'objet de nombreux travaux exploitant particulièrement la logique floue (Bordogna et Pasi 1991, Bordogna et al. 1991). L'approche générale est basée sur la définition de deux niveaux d'importance, l'un est lié aux critères de base et l'autre est lié à la stratégie d'agrégation de ces critères sans considérer toutefois les conditions qui peuvent les lier. A la différence de ces travaux, notre approche considère particulièrement le cas de requêtes comportant des préférences conditionnelles largement posées dans des cadres d'applications issues, à titre d'exemple, du domaine médical ou du domaine juridique. C'est, à notre connaissance, les premiers travaux sur l'utilisation des CP-Nets au problème de la formalisation de requêtes. Même si les évaluations expérimentales dans le cadre de ces applications ne sont pas menées à ce jour, nos travaux ont le mérite de poser un cadre formel pour l'expression et la quantification de l'importance des critères impliqués dans leur formulation.

– *Formalisation du besoin induit par la requête* : la spécificité de nos travaux, relativement aux autres travaux du domaine est sans doute l'intégration du profil de la session à travers l'historique de recherches de l'utilisateur, dans la phase de classification de la requête. Un prototype de classifieur basé sur l'approche proposée a été mis en œuvre puis entraîné et évalué sur des sous ensembles disjoint de requêtes issues des collections TREC 1, 10 et 9 comportant respectivement des requêtes informationnelles, navigationnelles et transactionnelles. Les précisions de la classification ont été comparées à un

classifieur classique, en l'occurrence *Timbl*, qui est basé seulement sur la morphologie de la requête et utilisé dans l'évaluation de l'approche de référence présentée dans (Kang et Kim 2003). Les résultats (Tamine et al. 2008b) montrent que notre approche est l'origine d'un accroissement de l'ordre de 30% pour les requêtes informationnelles. Les autres types de requêtes, notamment les requêtes transactionnelles, ne sont pas toutefois correctement détéctées pour raison de non exhautivité des descripteurs formalisés à ce jour. Une extension de ces descripteurs avec d'autres facteurs fera l'objet de nos travaux futurs.

3.6.2 Structuration et support de nos travaux

Nos travaux autour de la clarification du besoin en information de l'utilisateur se sont tout d'abord déroulés à la suite des mes travaux de thèse (Tamine 2000), comme une extension de mes propositions visant la définition de modèles d'AG basées sur des techniques avancées de l'algorithmique génétique. Ils se sont poursuivis à travers le co-encadrement de la thèse de *Fatiha Boubekeur* qui a abordé la problématique de formalisation et évaluation de requêtes préférentielles. En parallèle à ces travaux, et dans l'objectif de définir des cycles de vie appropriés à des requêtes de différents types (transactionnel, navigationnel et informationnel) dans le cadre de la thèse de *Mariam Daoud* que je co-encadre, nous nous sommes posés la problématique de la classification de la requête selon son type "latent"> et avons fait des propositions de stratégies basées sur la classification contextuelle. Ces dernières stratégie ont été évaluées sur de larges corpus de *TREC* dans le cadre du stage de master de *Dinh Ba Duy*.

3.7 Conclusion et perspectives

Les travaux présentés dans ce chapitre donnent un aperçu de nos travaux portant sur la clarification du besoin en information de l'utilisateur. Nos contributions portent sur divers aspects :

- Définition d'un mécanisme de reformulation de requête par réinjection de pertinence exploitant les atouts des AG's. Nous avons particulièrement fait usage de leurs propriétés de parallélisme implicite, de capacité de préservation des briques élémentaires, des transformations génétiques spécifiques et de la technique de nichage et spéciation en vue d'asseoir une recherche coopérative menée par une population de requêtes en vue d'atteindre différentes régions de pertinence. Ceci rejoint, à juste titre, le principe de polyreprésentation de la requête, prouvée comme étant une stratégie intéressante en RI (Ingwersen 1994b).

- Formalisation de requêtes comportant des préférences conditionnelles qualitatives. Dans ce cadre, nous avons particulièrement fait usage des propriétés des CP-Nets pour la représentation compacte des conditions et évaluation formelle considérant la propriété de

dominance.

– Détection du besoin en information induit par une requête dans le cadre du *web*. La détection est basée sur la combinaison des sources d'évidence issues à la fois de la structure et du contexte de recherche conduisant à une classification probabiliste.

Les perspectives de recherche ouvertes par ces travaux portent essentiellement sur deux points :

– La caractérisation de différents cycles de vie d'une requête depuis sa clarification puis formalisation, jusqu'à la présentation des résultats en passant par son évaluation et ce, en fonction d'une typologie d'intentions de recherche. En continuité avec nos travaux présentés dans ce chapitre, cette caractérisation portera tout d'abord sur les trois types d'intention véhiculées par une requête informationnelle, une requête navigationnelle et une requête transactionnelle. Il convient alors de cibler les facteurs du contexte à exploiter dans chacune des trois configurations puis formaliser les fonctions d'évaluation de la pertinence appropriées. Ce travail, concernant la requête informationnelle, est déjà en cours dans le cadre des travaux de thèse de *Mariam Daoud*. Les centres d'intérêt de l'utilisateur sont particulièrement exploités pour réordonnancer les résultats de recherche. Nous envisageons d'investir les autres types d'intentions véhiculées par les requêtes transactionnelles et requêtes navigationnelles. C'est en partie les objectifs poursuivis dans le cadre de la tâche 2.6 du projet *Quaero* auquel nous participons.

– L'amélioration de la technique de pondération automatique des requêtes basée sur les U-CPnets, en particulier la traduction des ordres de préférences qualitatifs en valeurs d'utilités correspondantes. L'approche proposée suggère des ordres de préférences uniformément distribués sur un domaine de valeurs données selon l'ordre de préférence qui y est spécifié. Il est alors impossible de prendre en compte des énoncés préférentiels modulés par des opérateurs linguistiques "extrêmes" (à l'exemple de : "je préfère de loin le jus d'orange au jus de pomme"). Pour pouvoir moduler ainsi les préférences utilisateur et en tenir compte lors de la pondération, les ordres de préférences qualitatifs devraient être traduits en ensembles flous de valeurs d'utilités. Ceci permettrait de *fuzzyfier* le langage de requête et fournirait le moyen pour une plus large expressivité des requêtes utilisateur.

Evaluation contextuelle des requêtes

4

4.1 Introduction

L'évaluation de requêtes traduit l'étape de traitement succédant éventuellement à celle de la clarification, présentée dans le chapitre 3, et qui consiste à l'apparier avec les documents en vue de sélectionner ceux qui sont potentiellement pertinents pour l'utilisateur. Plus précisément, l'évaluation dite contextuelle, a pour valeur ajoutée, la considération de l'évidence issue du contexte de l'utilisateur qui a émis la requête en question, en vue d'améliorer la précision de cet appariement et par conséquent améliorer la pertinence des résultats qui en sont issus. Cette évidence peut se décliner sous différents aspects, développés dans le chapitre 1, décrivant le contexte de recherche : les préférences de l'utilisateur, ses centres d'intérêt, son historique de recherche, son lieu etc. Pour notre part, en continuité avec nos travaux sur la formalisation de requêtes préférentielles d'une part et la définition de contextes cognitifs d'autre part, nous nous sommes particulièrement intéressés à (1) l'évaluation de requêtes guidée par les préférences de l'utilisateur vues comme des critères de recherche (2) l'évaluation de requêtes guidée par les centres d'intérêt de l'utilisateur.
Ce chapitre est organisé comme suit. La section 4.2 présente une synthèse des travaux portant sur l'évaluation contextuelle des requêtes, on y mettra l'accent sur le traitement des préférences et centres d'intérêt de l'utilisateur lors du processus d'appariement requête-document. La section 4.3 présente notre modèle sémantique pour le traitement des requêtes préférentielles. La section 4.4 développe notre modèle inférentiel permettant d'asseoir une évaluation de requêtes intégrant les centres d'intérêt de l'utilisateur comme composante formelle du modèle. La section 4.5 résume notre contribution à ce domaine de recherche. Enfin, la section 4.6 conclut le chapitre et annonce les perspectives envisageables.

4.2 Synthèse des travaux du domaine

Sous un angle très large, l'accès contextuel à l'information est fondé sur trois paradigmes : filtrage et/ou recommandation, navigation, accès par requête.

– *Filtrage et/ou recommandation* : la recommandation fut parmi les premières approches proposées dans le but de fournir un accès contextuel à l'information. Un modèle de recommandation est généralement basé sur un modèle d'utilisateur individuel impliquant un filtrage d'information par le contenu (ou cognitif) (Lieberman 1995, Pazzani et al. 1996, Boughanem et al. 2002b) ou groupe d'utilisateurs ciblant un filtrage d'information collaboratif (ou social) (Resnick et al. 1994, Mobasher et al. 2000, Balabanović et Shoham 1997). Le filtrage/recommandation repose sur une démarche proactive, initiée par le système, qui a pour objectif de recommander à l'utilisateur ou à son groupe, des informations selon son modèle appris implicitement ou explicitement. L'accès est qualifié précisément de *personnalisé* puisque l'aspect contextuel est couvert par l'utilisateur lui même.

– *Navigation* : l'accès contextuel se base dans ce cas sur la structure hypertexte du réseau d'information. Dans ce sens, la plupart des travaux ont proposé une variante personnalisée de l'algorithme *PageRank* en l'occurrence l'algorithme PROS (Chirita et al. 2004), Hubrank (Jeh et Widom 2002), *topic sensitive rank* (Haveliwala 2002b). L'idée fondamentale y est de recalculer les valeurs d'autorité des pages selon leur degré de "proximité" aux centres d'intérêt de l'utilisateur.

– *Accès par requête* : c'est précisément l'accès classique déployé dans les SRI et qui fait l'objet de nos questions de recherche. L'accès contextuel couvre dans ce cas, le processus d'appariement entre une requête dûment exprimée par l'utilisateur et une collection de documents en vue de sélectionner ceux qui sont pertinents en considérant le contexte de recherche en cours.

  Dans ce sens, les travaux que nous avons menés, comme retracés dans le chapitre introductif, ont porté sur un contexte "plat" relevant d'une approche orientée-système, mettant quelque peu en marge le rôle de l'utilisateur dans l'activité de recherche d'information, puis un contexte "profond" de par la diversité des dimensions qu'il comprend, remettant l'utilisateur au centre de l'activité de recherche d'information. Dans ce sens, nous nous intéresserons dans la suite à des travaux connexes portant sur :

  1. *l'évaluation flexible de requêtes* mettant en œuvre des requêtes préférentielles, telles que formalisées dans le chapitre 3, section 3.4. Le contexte est réduit à la simple expression de critères de recherche. Ces critères (inhérents à la requête) sont évalués de manière analogue pour des utilisateurs différents ayant exprimé la même requête.

  2. *l'évaluation personnalisée de requêtes* mettant en œuvre un appariement requête-document-contexte cognitif de l'utilisateur. Ce dernier comprend les centres d'intérêt de l'utilisateur tels que

spécifiés dans le chapitre 2, section 2.3. Dans ce cas, une même requête émanant de deux utilisateurs différents donne lieu à des réponses différentes, ce qui évoque proprement l'évaluation personnalisée de la requête.
Notons de prime abord, qu'on ne relève pas actuellement dans la littérature, une différence particulière avec l'évaluation contextuelle, si ce n'est la locution utilisée. Toutefois, il est à noter que l'expression "accès ou évaluation personnalisé" a germé en premier (années 1990) avec l'apparition des systèmes de recommandation. L'expression "évaluation contextuelle" est apparue plus tard (années 2000) pour couvrir divers aspects du contexte (Cf. chapitre 1, section 1.3). Les deux expressions traduisent l'exploitation d'un ou plusieurs facteurs du contexte de recherche pour délivrer l'information appropriée à l'utilisateur. Pour notre part, le facteur est représenté par les centres d'intérêt de l'utilisateur ; c'est à juste titre le facteur le plus largement exploité par les travaux du domaine.

4.2.1 Evaluation de requêtes préférentielles

L'évaluation de requêtes préférentielles dans le domaine de la RI, traduit la prise en compte des préférences exprimées dans la requête dans le processus de sélection des documents pertinents. En pratique, ceci se traduit par la définition d'une fonction d'évaluation de la pertinence (RSV[1]) dont la sémantique dépend des conditions décrites par l'expression des préférences. Les approches classiques d'évaluation procèdent d'abord à l'évaluation de chaque terme pondéré de la requête (second niveau de préférences) indépendamment des autres puis à leur agrégation selon les conditions d'agrégation définies dans la requête (premier niveau de préférences) (Cater et Kraft 1989).
L'évaluation individuelle des critères de recherche dépend de la sémantique du poids associé à chaque critère de recherche. A titre d'exemple, selon la sémantique du seuil (Cf chapitre 3, section 3.2), l'approche d'évaluation consiste à favoriser les termes avec une valeur de pertinence g supérieure à leur poids a considérée ici comme un seuil. La fonction d'évaluation est définie (Buell et Kraft 1981b) :

$$g(F(d,t),a) = \begin{cases} P(a) \times \frac{F(d,t)}{a} \; si \; F(d,t) < a \\ P(a) + Q(a) \times \frac{F(d,t)-a}{1-a} \; sinon \end{cases} \tag{4.1}$$

où P(a), Q(a) sont utilisés pour ajuster le comportement du seuil : $P(a) = \frac{1+a}{2}$,$Q(a) = \frac{1-a^2}{4}$. Ainsi pour $F < a$, la fonction g évalue la proximité de F avec le seuil a, et pour $F > a$, elle estime dans quelle mesure la valeur dépasse, c'est à dire est plus satisfaisante que le seuil a.
D'autres formulations ont été également proposées pour la sémantique d'importance relative (Dubois et Prade 1986) et sémantique de la perfection (Bordogna et al. 1991).

[1]Relevance Status Value

L'évaluation globale dépend, rappelons le, de la structure de la requête. Dans le cas particulier des requêtes conjonctives et requêtes disjonctives, l'évaluation se fait à l'aide des t-normes et t-conormes respectivement (Dubois et Prade 1985). On peut citer comme t-normes : $min(g_1, g_2)$, ou $g_1 \times g_2$ ou encore $max(g_1 + g_2 - 1, 0)$. Des exemples de t-conormes sont : $max(g_1, g_2)$ et $min(1, g_1 + g_2)$.
Dans le cas de requêtes plus flexibles exprimées à l'aide d'opérateurs linguistiques, d'autres approches d'évaluation basées sur les *OWA* ont été proposées (Bordogna et Pasi 1991, Boughanem et al. 2005).

Notre contribution dans ce cadre porte particulièrement sur l'évaluation des requêtes préférentielles conditionnelles. En accord avec le formalisme adopté (présenté dans le chapitre 3, section 3.4), notre approche d'évaluation est fondée sur un modèle d'appariement de graphes CP-Nets, présenté en section 4.3.

4.2.2 Evaluation personnalisée de requêtes

L'évaluation personnalisée de requêtes traduit l'exploitation d'un ou de plusieurs facteurs du contexte comme source d'évidence pour fournir des réponses appropriées à l'utilisateur. Comme évoqué dans le chapitre introductif, la littérature révèle l'exploitation effective du lieu et du temps dans le cadre de la recherche contextuelle dans des environnements mobiles (Göker et al. 2004, Iqbal et al. 2005) ou de la tâche (Dumais et al. 2003a, Westerveld et al. 2001), du cadre applicatif (Lin et al. 2005a, Lee et al. 2005a, Cheverst et al. 2000) et des centres d'intérêt de l'utilisateur. Notre intérêt porte précisément sur l'exploitation du facteur *centres d'intérêt*. Les travaux dans ce sens ont pour objectif de réviser le procédé d'appariement classique requête-document soit *a priori*, en intégrant les centres d'intérêt comme composante formelle du modèle d'appariement, soit *a posteriori* en utilisant l'évidence issue des centres d'intérêt de l'utilisateur en vue de réordonner les résultats issus d'un appariement classique requête-document.
A notre connaissance, comme également rapporté dans (Micarelli et al. 2007), peu de travaux ont porté sur la modélisation d'un appariement requête-document-centres d'intérêt (Fan et al. 2004, Lin et al. 2005b). Dans (Fan et al. 2004), les auteurs proposent d'adapter les paramètres de la fonction de pertinence au contexte de l'utilisateur, en utilisant les techniques de programmation génétique. Dans (Lin et al. 2005b), les auteurs utilisent le modèle PLSA (Hofmann 1999) pour déterminer une intention de recherche latente à partir de la requête et de l'utilisateur décrit par ses centres d'intérêt. Suivant une autre approche, la modélisation de l'appariement requête-document-centres d'intérêt est abordée dans (Agichtein et al. 2006, Teevan et Dumais 2005) comme un problème de prédiction des préférences des utilisateurs, en termes de documents, en utilisant des stratégies d'apprentissage automatique. Les centres d'intérêt sont à court terme, inférés à partir du comportement de l'utilisateur lors du processus de navigation.
La technique de réordonnancement des résultats de recherche est la plus utilisée dans le domaine (Sieg et al. 2007, Gauch et al. 2003, Liu et Yu

2004, Bai et al. 2007). Elle est généralement basée sur une combinaison linéaire monovaluée (Sieg et al. 2007, Gauch et al. 2003, Bai et al. 2007, Daoud et al. 2008c) ou multivaluée (Liu et Yu 2004) des scores de pertinence requête-document et scores de pertinence requête-centres d'intérêt. Pour notre part, nous avons abordé le problème de l'évaluation personnalisée de requêtes en proposant des solutions basées sur la spécification d'un modèle formel d'accès à l'information (Tamine et al. 2008a; 2007b, Tamine et Boughanem 2006), et d'autres basées sur le réordonnancement des résultats de recherche (Daoud et al. 2008c;a). Cette dernière contribution, suit la lignée des travaux inscrits dans cette approche, en utilisant la technique de combinaison des scores de pertinence. Nous estimons que notre forte contribution dans le cadre de l'évaluation personnalisée de requête se traduit par la spécification d'un modèle formel, à l'instar du modèle décrit dans (Lin et al. 2005b), qui intègre les centres d'intérêt de l'utilisateur comme composante à part entière du modèle. C'est précisément cette contribution au domaine qui est étayée dans la section 4.4.

4.3 Modèle flexible pour l'évaluation des requêtes préférentielles

Le but de l'évaluation d'une requête préférentielle est de retourner à l'utilisateur des documents qui satisfont les contraintes imposées par les préférences qui la caractérisent. Pour notre part, en accord avec le formalisme de représentation des requêtes à base de CP-Nets, présenté dans le chapitre 3, nous proposons une approche d'évaluation permettant d'apparier des requêtes CP-Nets avec les représentations des documents. Dans un souci d'analogie avec les CP-Nets, les représentations des documents ont été adaptées nous permettant à terme de faire asseoir un modèle d'évaluation orienté CP-Nets. A ce titre, deux types de représentations des documents ont été proposées conduisant ainsi à deux modèles d'évaluation de requêtes : modèle d'évaluation basé sur l'agrégation floue et modèle d'évaluation basé sur l'appariement de graphes CP-Nets. Les descriptions de ces deux modèles sont synthétisées dans ce qui suit.

4.3.1 Evaluation basée sur l'agrégation floue

Notre approche d'évaluation est basée sur les étapes suivantes (Boubekeur et al. 2006) :

1. le processus de recherche est lancé dans un premier temps sur l'ensemble des termes de la requête CP-Net sans tenir compte de la pondération au préalable. Le résultat est une liste de documents pertinents probables pour la requête,
2. les documents retrouvés sont ensuite représentés par des CP-Nets, puis documents et requêtes sont reformulés en expressions booléennes,

3. un processus d'évaluation calcule la valeur de pertinence de tels documents pour la requête UCP-Net, et ordonne les documents par degré de pertinence.

La première étape est une recherche classique. Les documents qui apparient les termes de la requête sont alors retournés par le système. Les étapes 2 et 3 sont propres à notre approche, nous les décrivons dans ce qui suit.

A. Représentation CP-Net du document

L'objectif fondamentale de cette étape est de produire une représentation du document qui s'apparie avec la requête en terme de (1) structuration des variables représentatives (2) quantification de ces variables relativement au contenu du document. A cet effet, chaque document supposé pertinent pour une requête $Q = (V, E)$ est représenté par un CP-Net $D = (V, E')$ dans le même espace de termes que la requête. La topologie correspondante est semblable à celle du CP-Net requête Q mais les tables CPT sont différentes. En effet, les CPT dans le CP-Net document D quantifient numériquement l'importance des termes d'indexation dans D. Cette importance se traduit par les poids des termes correspondants dans D. Les poids sont généralement exprimés par une variante de $tf \times idf$. Le document (respectivement la requête) est alors traduit en expression booléenne, comme une disjonction de conjonctions. Chaque conjonction étant construite sur l'ensemble des éléments du produit cartésien $Dom(X_1) \times Dom(X_2) \times \ldots \times Dom(X_n)$ où les X_i $(1 \leq i \leq n)$ sont les nœuds du CP-Net document (respectivement CP-Net requête). Formellement, on a (Boubekeur et al. 2006) :

$$D = \vee_{j_i}(\wedge_i(t_{i,j_i}, p_{i,j_i}) \tag{4.2}$$

$$Q = \vee_{j_i}(\wedge_i(t_{i,j_i}, f_{i,j_i}) \tag{4.3}$$

où $t_{i,j_i} \in Dom(X_i)$, $1 \leq i \leq n$, $1 \leq j_i \leq |Dom(X_i)|$, p_{i,j_i} est le poids de t_{i,j_i} dans D (basé sur sa fréquence d'occurrence), f_{i,j_i} est le poids du terme t_{i,j_i} dans la requête Q étant donnée une valeur de ses parents.

Soit $m = |Dom(X_1)| \times |Dom(X_2)| \times \ldots \times |Dom(X_n)|$, en posant $\wedge_i(t_{i,j_i}) = T_k$, $1 \leq k \leq m$ les représentations (4.2) et (4.3) sont respectivement réduites à :

$$D = \vee_k(T_k, S_k) = \vee(T_k, S_k) \tag{4.4}$$

$$Q = \vee_k(T_k, U_k) = \vee(T_k, U_k) \tag{4.5}$$

où S_k et U_k sont les poids agrégés des valeurs f_{i,j_i} et p_{i,j_i} respectivement, calculés en respectant le principe d'agrégation additive.

B. Evaluation de la requête

L'objectif fondamentale de cette étape est de calculer la valeur de pertinence $RSV(Q, D)$ de chaque document D pour la requête Q en cours

d'évaluation. Formellement, soit Q un CP-Net requête exprimée sous forme d'une expression booléenne en forme normale disjonctive exprimée selon la formule 4.4, et D le document retourné est exprimé sous forme booléenne tel que défini dans la formule 4.5. Nous proposons d'adapter et d'utiliser l'opérateur du minimum pondéré comme suit :
Soit U_K le poids d'importance de T_k dans Q, $F(D, T_k) = S_k$ le poids de T_k dans le document D. On note $RSV_{T_k}(U_k, F(D, T_k))$ la fonction d'évaluation de T_k pour le document D. Les différentes conjonctions pondérées (T_k, U_k) étant liées par une une disjonction, ce qui donne :

$$RSV_{T_k}(U_k, F(D, T_k)) = min(U_k, S_k) \tag{4.6}$$

$RSV(Q, D)$ est alors obtenue par agrégation de l'ensemble des poids de pertinence calculés dans 4.9 comme suit :

$$RSV(Q, D) = max_k(min(U_k, S_k)) \tag{4.7}$$

4.3.2 Evaluation basée sur l'appariement de graphes CP-Nets

A la différence de l'approche d'évaluation de requête précédente, les documents sont représentés, dans cette approche, selon le formalisme CP-Nets indépendamment des requêtes en cours d'évaluation. Cette approche d'évaluation est fondée sur deux points clés (Boubekeur et al. 2007; 2008) :

1. la représentation sémantique des documents permettent de mettre en évidence les concepts représentatifs de leur contenu d'une part et les relations entre ces concepts d'autre part. Cette représentation est exprimée à l'aide d'un graphe orienté acyclique par analogie aux CP-Nets,
2. l'évaluation de requête par appariement des graphes CP-Nets requête et document.

Nous donnons dans ce qui suit un bref aperçu des stratégies de représentation et évaluation sous-jacentes à cette approche.

A. Représentation sémantique des documents

Le processus de représentation sémantique des documents s'articule principalement autour de trois étapes. La première étape concerne l'identification des concepts représentatifs des documents, la seconde étape concerne la découverte des relations entre ces concepts et la troisième étape porte sur l'intégration des concepts et des relations associées au sein d'un formalisme unifié, en l'occurrence un graphe CP-Net. Ces étapes, illustrées sur la figure 4.1, sont synthétisées comme suit :

1. ***Identification des concepts***
   Les concepts sont extraits à partir des termes représentatifs du contenu sémantique du document, par projection sur l'ontologie générale *WordNet*. Lors de cette projection, si plusieurs concepts correspondent à un terme donné, le terme est désambiguïsé. Les sous étapes de cette première étape sont :

- *l'identification des termes* : le but de cette étape est d'identifier des mono ou multi termes dans le document. Ces termes correspondent à des entrées dans l'ontologie,
- *la pondération des termes* : dans cette étape, on propose une variante de $tf \times idf$, s'appliquant aux mono et aux multi-termes. Le but est d'éliminer les termes les moins fréquents dans le document et de maintenir seulement les termes les plus représentatifs,
- *la désambiguïsation* : les termes d'index sont associés à des sens (*synsets*) correspondants dans l'ontologie. Chaque terme extrait pouvant avoir plusieurs sens possibles, le but de cette étape est de sélectionner le meilleur sens du terme dans le document.

2. ***Découverte de relations entre concepts***
   Les relations contextuelles entre les concepts extraits sont découvertes en utilisant une approche que nous proposons, basée sur la technique des règles d'association,

3. ***Construction du CP-Net document***
   les concepts et les relations correspondantes sont organisés en un graphe conceptuel, à savoir le graphe CP-Net. Les nœuds du CP-Net sont les concepts représentatifs du document. Les arcs du CP-Net traduisent les relations entre concepts.

Une description détaillée et illustrée de ce processus de représentation sémantique des documents est présentée dans (Boubekeur 2008).

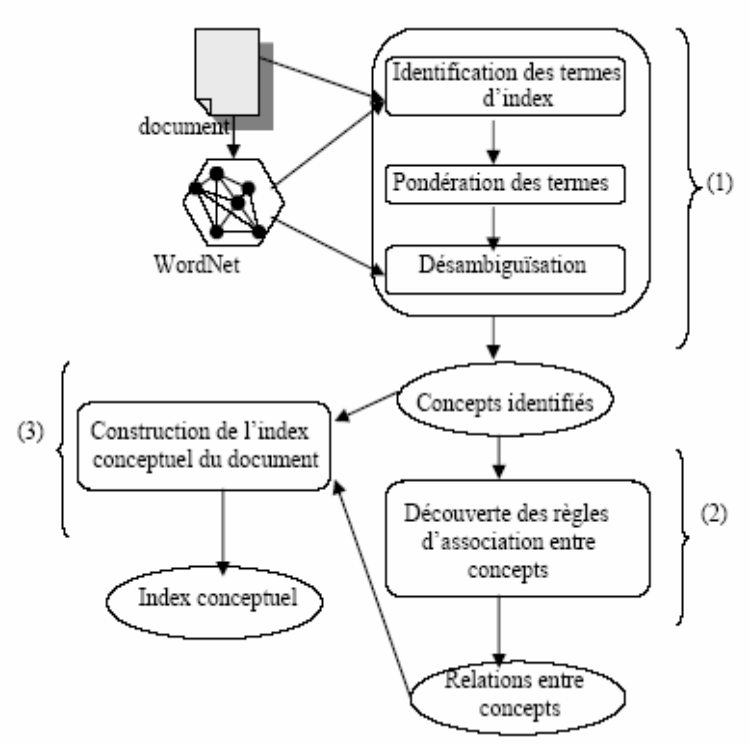


FIG. 4.1 – *Construction de l'index conceptuel du document*

B. Evaluation de requête par appariement de graphes CP-Nets

Comparativement à notre proposition présentée dans le paragraphe 4.3.1, plutôt que d'interpréter les CP-Nets document et requête en expressions booléennes pour évaluer leur degré de correspondance, nous proposons ici une mesure de similarité issue de l'appariement des graphes CP-Net document et requête. Cette similarité est calculée comme agrégation des similarités partielles des deux graphes à travers leurs concepts communs exprimée comme suit (Boubekeur et al. 2007) :

$$SIM(G_D, G_Q) = \frac{|\eta(D) \cap \eta(Q)|}{|\eta(D) \cup \eta(Q)|} \times max(Sim^X_{X \in \eta(D) \cap \eta(Q)}(D, Q)) \qquad (4.8)$$

où G_D et G_Q sont les graphes CP-Nets correspondant respectivement au document D et à la requête Q. $Sim^X(D, Q)$ est la similarité partielle entre le document D et la requête Q au niveau du concept X. En se basant sur la topologie des graphes CP-Nets, cette mesure est calculée comme combinaison de la similarité structurelle et de la similarité relationnelle comme suit :

$$Sim^X(D, Q) = \alpha \times Sim^X_{struct}(D, Q) + (1 - \alpha) \times Sim^X_{relat}(D, Q) \qquad (4.9)$$

où α, $0 \leq \alpha \leq 1$ est une valeur donnée qui spécifie l'importance de la similarité structurelle par rapport à la similarité relationnelle.
La similarité structurelle Sim^X_{struct} définit la proportion de valeurs (instances) de X communes dans le document D et requête Q. Une instance commune dans D et Q est un terme de la requête Q qui appartient au document D.
La similarité relationnelle Sim^X_{relat} indique le degré de représentativité du concept (nœud) X correspondant à son importance aussi bien dans le document que dans la requête. Sim^X_{relat} est mesurée en fonction de la profondeur associée au concept X dans la hiérarchie correspondant au graphe CP-Net. Les définitions formelles des similarités relationnelle et structurelle sont données dans ce qui suit :

- **Similarité structurelle**

$$Sim^X_{struct}(D, Q) = \frac{|Dom_{X,D} \cap Dom_{X,Q}|}{|Dom_{X,D} \cup Dom_{X,Q}|} \qquad (4.10)$$

  où $Dom_{X,D}$ et $Dom_{X,Q}$ sont les domaines des instances (valeurs) associées au concept-nœud X respectivement dans G_D et G_Q.

- **Similarité relationnelle**

$$Sim^X_{relat}(D, Q) = \sum_j Sim^{X_j}_{relat}(D, Q) \qquad (4.11)$$

  où $Sim^{X_j}_{relat}(D, Q)$ définit la similarité relationnelle de D à Q au niveau de l'instance X_j de X. Cette valeur est calculée comme le minimum entre son degré d'importance dans le document et dans la requête, normalisé par la somme de ses degrés d'importance dans le document et la requête CP-Net. Le degré d'importance de la valeur

X_j de X est défini comme le produit de son poids de représentativité dans le document ou dans la requête, et de son degré d'importance (défini par rapport à la position du nœud correspondant dans le graphe CP-Net) dans le CP-Net correspondant. Formellement :

$$Sim_{relat}^{X_j}(D,Q) = \frac{min(W_{X_j,D} \times Deg_D(X), W_{X_j,Q} \times Deg_Q(X))}{W_{X_j,D} \times Deg_D(X) + W_{X_j,Q} \times Deg_Q(X)} \quad (4.12)$$

où $W_{X_j,D}$ et $W_{X_j,Q}$ sont les poids associés aux valeurs X_j de X respectivement dans le document D et requête Q, $Deg_D(X)$ et $Deg_Q(X)$ représentent le niveau d'importance de X respectivement dans le document D et la requête Q. Le niveau d'importance du concept-nœud X est inversement proportionnel à la profondeur du nœud correspondant dans le graphe. Ainsi, pour un graphe de profondeur maximale n, la racine du graphe est de niveau 1 et d'importance 1. Ses descendants directs sont de niveau 2 et d'importance $1/2$ etc. Les éléments de niveau n ont une importance de $1/n$.

4.4 Modèle inférentiel pour l'évaluation personnalisée des requêtes

4.4.1 Motivations

La revue de l'état de l'art portant sur la personnalisation du processus d'évaluation de la requête révèle une abondance des travaux proposant des heuristiques de réordonnancement des résultats de recherche issus d'un appariement classique requête-document. La qualité de ce réordonnancement dépend en grande partie de la qualité de l'ordre initial des résultats d'une part et de la qualité du modèle de l'utilisateur d'autre part. Les travaux inscrits dans cette direction sont encore récents (depuis 2003), en cours d'évolution, il est alors difficile à ce jour de dresser un bilan critique. La direction que nous avons entreprise est différente, en accord avec notre ambition de spécifier un modèle formel dédié à l'accès personnalisé à l'information, se démarquant de simples techniques de réordonnancement qui augmentent un modèle d'accès classique. Notre objectif est simplement le suivant : *plutôt que de réviser la valeur de pertinence d'un document en considérant a posteriori les centres d'intérêt, considérer ces derniers a priori comme faisant partie du modèle d'accès au même titre que les requêtes et les documents dans un modèle classique.* Ceci rejoint l'idée de substituer la fonction de pertinence classique qui mesure le degré d'appariement requête-document $RSV(Q,D) = p(Q/D)$, par une fonction indexée par l'utilisateur $RSV_u(Q,D) = p(D/Q,U)$ où $p(A/B)$ est la probabilité conditionnelle de l'événement A sachant l'événement B et U représente l'utilisateur décrit par ses centres d'intérêt. Plus particulièrement, nous abordons l'évaluation de la pertinence d'un document comme un problème décisionnel lié à la présentation ou non d'un document en fonction de plusieurs critères :

- la pertinence du document relativement à la requête,
- la pertinence du document relativement aux centres d'intérêt de l'utilisateur,
- et l'adéquation des centres d'intérêt de l'utilisateur vis-à-vis de la requête.

Dans ce sens, nous formalisons ce problème à l'aide d'un modèle inférentiel, basé sur les diagrammes d'influence (DI) (Shachter 1988), extension des réseaux Bayésiens (RB) (Jensen 2001). Les DI constituent en effet un cadre théorique approprié pour la formalisation du problème décisionnel lié à la présentation d'un document à l'utilisateur compte tenu de son utilité liée à l'influence mutuelle de son contenu, de la requête émise et des centres d'intérêt spécifiques de l'utilisateur.

4.4.2 Formalisation du problème

Intuitiuvement, l'évaluation personnalisée d'une requête se traduit par l'assertion suivante :
Etant donné une requête Q, l'objectif du SRI est d'identifier les documents D_j qui sont appropriés au besoin en information de l'utilisateur U.

Nous déclinons ce problème par la maximisation d'une probabilité conditionnelle de pertinence $p(d/q,u)$ dépendant à la fois de la requête q et de l'utilisateur u, où d_j, q et u sont les variables aléatoires associées à D_j, Q et U respectivement . Nous développons dans ce qui suit, les étapes fondamentales de cette formalisation (Tamine et al. 2007b) :

La probabilité $p(d_j/q,u)$ peut être formulée en appliquant la loi de Bayes comme suit :

$$p(d_j/q,u) = \frac{p(q/d_j,u)p(d_j/u)}{p(q/u)} \tag{4.13}$$

Etant donné que le dénominateur $p(q/u)$ est constant pour une requête et un utilisateur donnés, le numérateur suffit pour ordonner les documents selon leur pertinence. Ainsi la fonction d'ordre RSV (*Relevance Status Value*) d'un document peut être définie comme suit :

$$RSV_u(Q,D_j) = p(q/d_j,u)p(d_j/u) \tag{4.14}$$

On constate que le premier membre de cette équation ($p(q/d_j,u)$), dépendant de la requête, traduit le degré de satisfaction de la requête Q en tenant compte du document D_j et de l'utilisateur U. Le second membre, indépendant de la requête, exprime en revanche la probabilité de pertinence du document D_j en tenant compte de l'utilisateur U. Ainsi, si on considère que l'utilisateur U est représenté par ses centres d'intérêt multiples (de nombre n), modélisés par $C_1, C_2, ..., C_n$, alors l'équation 4.14 du RSV devient :

$$RSV_u(Q,D_j) = p(q/d_j,c_1,c_2,...,c_n)p(d/c_1,c_2,...,c_n) \tag{4.15}$$

où c_k représente la variable aléatoire associée au centre d'intérêt C_k de l'utilisateur. On relève deux points importants :

1. Deux conditions sont nécessaires quant à la décision de sélection d'un document :
   - *pertinence thématique* exprimant l'adéquation du contenu du document au besoin formulé par la requête,
   - *condition d'utilité* exprimant le degré d'adéquation du document D_j à l'ensemble des centres d'intérêt de l'utilisateur.
2. La couverture thématique d'un document au besoin en information exprimé par l'utilisateur à travers sa requête, est réalisée en maximisant la corrélation de l'information à ses différents centres d'intérêt.

De notre point de vue, l'objectif du SRI peut être affiné vers l'identification de documents pertinents pour la requête et utiles vis-à-vis des centres d'intérêt de l'utilisateur. Cette problématique peut être formalisée globalement à travers le diagramme d'influence, noté $ID(D_v, C_v, \mu)$, spécifié à travers :

- l'ensemble des variables $D_v = \{d_1, d_2, ..., d_m\}$ représentant les documents , où n est le nombre total de documents dans la collection,
- l'ensemble des variables $C_v = \{c_1, c_2, ..., c_n\}$ représentant les centres d'intérêt de l'utilisateur, où c_n est le n^{ime} centre d'intérêt considéré de la librairie des centres d'intérêt de son profil,
- l'ensemble des utilités $\mu = \{\mu_1, \mu_2, ..., \mu_n\}$, où μ_k exprime l'utilité du document instancié, D pour le centre d'intérêt C_k de l'utilisateur.

Ainsi, l'objectif revient à ordonner les documents $d_j \in D_v$ selon la valeur de l'utilité exprimée par $\mu(d_j) = \Psi(\mu_1, \mu_2, ..., \mu_n)$, où Ψ est un opérateur approprié d'agrégation combinant les valeurs des évidences issues de l'ensemble $c_1, c_2, ..., c_n$. En considérant l'équation 4.14, le calcul du score de pertinence est exprimé par :

$$RSV_U(Q, D_j) = \Psi_{k=1..n}(\mu_k(d_j, c_k) * p(q/d_j, c_1, c_2, ..., c_n)) \tag{4.16}$$

4.4.3 Topologie du modèle

La topologie du modèle d'accès est représentée dans la figure 4.2 par le graphe acyclique $G = (V, E)$ où V comprend les nœuds représentant des variables aléatoires X_1, X_2, ..., X_n. A chaque variable X_i est associée un ensemble de valeurs mutuellement exclusives définies dans $dom(X_i)$. L'ensemble E comprend les arcs existants entre les nœuds qui traduisent des relations de causalité décrites par des probabilités conditionnelles attachées à chaque nœud.

La topologie met en évidence trois principaux nœuds :

1. Nœuds chance : représentant les différents types d'informations propagées dans le diagramme : à savoir la requête, le document et le centre d'intérêt,
2. Nœuds décision : représentent les variables de décision du modèle, attachées dans notre cas aux documents

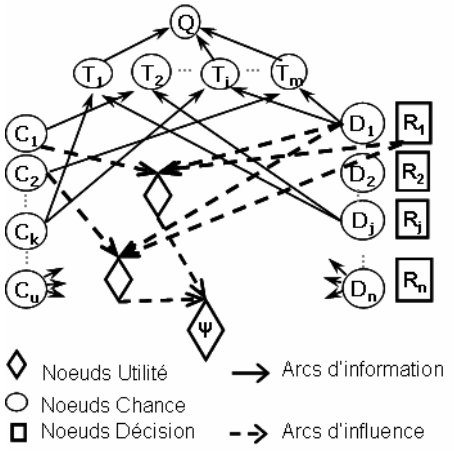


FIG. 4.2 – *Topologie du modèle d'évaluation personnalisée de la requête*

3. Nœuds utilité : chaque nœud exprime *l'utilité de la décision de présenter un document compte tenu des centres d'intérêt* de l'utilisateur.

Une description détaillée de la topologie du diagramme est présentée dans (Tamine et al. 2007b).

4.4.4 Principe de l'évaluation de requêtes

Le processus d'évaluation de requête a pour objectif d'assigner à chaque document un score de pertinence relative aux centres d'intérêt de l'utilisateur qui l'a émise. Ce score est calculé, dans le cas de notre approche en maximisant son utilité globale. Nous avons proposé différentes formalisations de la fonction de pertinence dont :

$$RSV_U : \begin{cases} R \longrightarrow R \\ RSV_U(Q, D_j) \mapsto EU(r_j/D_j) \end{cases} \tag{4.17}$$

où $EU(r_j/D_j)$ est l'utilité globale associée à la décision "*D est pertinent, peut être présenté à l'utilisateur*" calculée comme suit :

$$EU(r_j/D_j) = \Psi_{k=1..n} \left[\mu_k(r_j/d_j, c_k) * p(q/d_j, c_k) \right] \tag{4.18}$$

Un seul document est instancié positivement ($D_j = d_j$) à la fois, il en est de même pour les centres d'intérêt C_k. La propagation de l'information est déclenchée par ces instanciations. La propagation dans ce modèle consiste alors à calculer, pour chaque nœud, la probabilité *a posteriori* étant donné les probabilités conditionnelles et marginales *a priori*. La propagation tente de calculer la probabilité que la requête soit satisfaite étant donné un document instancié à $D_j = d_j$ et un centre d'intérêt instancié à $C_k = c_k$. Pour chaque document, ce processus est réitéré pour tous les centres d'intérêt ; puis l'algorithme reprend pour considérer tous les documents de la collection candidats à l'évaluation.
Grâce au principe de marginalisation inhèrent aux modèles inférentiels (Jensen 2001), l'utilité globale d'un document pour une requête donnée est calculée comme :

$$EU(r_j/D_j) = \Psi_{k=1..n}\left[\mu_k(r_j/d_j, c_k) * \sum_{\theta^s \in \theta}(p(q/\theta^s) * \prod_{T_i \in Q \cap (D_j \cup c_k)} p(\theta_i^s/d_j) * p(\theta_i^s/c_k))\right] \quad (4.19)$$

où Ψ est un opérateur d'agrégation approprié qui sera spécifié dans la section 4.4.6.

4.4.5 Distribution de probabilités

La quantification des distributions de probabilités dans le modèle, consiste à donner une signification aux arcs reliant les différents types de nœuds du réseau, et à estimer l'utilité des décisions de présentation des documents pertinents compte tenu de tous les centres d'intérêt de l'utilisateur. Ainsi, la composante quantitative du modèle comprend des distributions de probabilités conditionnelles où pour chaque variable $x_i \in V$, est attachée une classe de probabilités $p(X) = p(x_i/pa(x_i))$ qui est fonction de toutes les configurations possibles de ses nœuds parents $pa(x_i)$ dans le réseau G, notée θ. L'estimation des distributions de probabilités pour quantifier chacun des membres de l'équation de l'utilité globale 4.19 est fondée sur les éléments suivants :

- *Estimation de la probabilité de la satisfaction de la requête* : cette estimation repose sur l'intuition que plus le nombre de termes instanciés positivement dans la configuration est élevé, plus le degré de satisfaction de la requête est important. Pour cela, nous avons appliqué des opérateurs d'agrégation flexible tels que le *Noisy-OR* qui est reconnu effeicace dans de nombreuses applications (Pearl 1988).

- *Estimation de la pertinence des termes dans les documents et centres d'intérêt* : nous avons utilisé des heuristiques classiques permettant de décliner l'importance d'un terme dans une unité d'information sur la base de sa distribution dans l'unité élémentaire elle même et dans la collection. A titre d'exemple, la pertinence probabilité de pertinence d'un terme T_i dans un document D_j est calculée comme suit :

$$p(t_i/d_j) = \begin{cases} \frac{wtd(T_i,D_j)}{\sum_{T_l \in \tau(D_j)} wtd(T_l,D_j)} \; if \; T_i \in \tau(D_j) \\ \delta_d \; sinon \end{cases} \quad (4.20)$$

  où d_j est la variable aléatoire associée au document D_j, t_i est la variable aléatoire associée au terme T_i (T_l, D_j) et $wtc(T_l, C_k)$ sont respectivement les poids du terme T_l dans le document D_j et centre d'intérêt C_k, δ_d et δ_c sont des constantes $(0 \leq \delta_d, \delta_c \leq 1)$ exprimant la probabilité par défaut liée à l'ignorance portée sur la pertinence d'un terme. Par conséquent, on calcule la probabilité de non pertinence d'un terme vis-à-vis d'un document comme suit : $p(\overline{t_i}/d_j) = 1 - p(t_i/d_j)$.

- *Estimation de l'utilité des décisions* : la valeur de l'utilité élémentaire exprime le degré de concordance entre le centre d'intérêt instancié et le document observé. Elle est liée à la décision de restituer

le document D_j lorsque l'on observe le centre d'intérêt C_k. Sa valeur est d'autant plus élevée que le centre d'intérêt est "proche" du document observé. Nous avons évalué de nombreuses métriques basées sur la similitude vectorielle ou ensembliste (Tamine et al. 2008a, Zemirli et al. 2007).

4.4.6 Opérateurs d'agrégation

Dans le but de mesurer l'utilité globale du document pour l'ensemble de tous les centres d'intérêt, on propose d'agréger les valeurs des utilités élémentaires correspondantes de manière à satisfaire les relations pouvant exister entre les centres d'intérêt de l'utilisateur selon les deux hypothèses suivantes (Tamine et al. 2007b) :

- ***Hypothèse 1*** : *les centres d'intérêt de l'utilisateur sont indépendants*
  Dans ce cas, l'impact du centre d'intérêt est considéré isolément dans le calcul de l'utilité globale. En effet, le score de pertinence du document devrait être plus important pour les centres d'intérêt proches de la requête que ceux qui ne le sont pas. En d'autres termes, étant donné l'indépendance des centres, leur influence sur le processus d'inférence est fortement guidée par le centre le plus important dans le contexte de la requête. Il est donc opportun de choisir l'opérateur d'agrégation basé sur le consentement relatif (Fargier et Perny 2003) permettant de considérer l'importance relative des centres d'intérêt dans la couverture du sujet de la requête. Une formulation possible de cet opérateur d'agrégation peut être donnée en maximisant les valeurs des utilités élémentaires comme suit :
  $$\Psi(\mu_1, ..., \mu_n) = Max\,(\mu_1, ..., \mu_n) \tag{4.21}$$
  où u est le nombre total des centres d'intérêt et $\mu_k = \mu_k(r_j/d_j, c_k) * p(q/d_j, c_k)$ l'utilité élémentaire d'ordre k associée à la décision de restituer le document D_j en considérant le centre d'intérêt C_k.

- ***Hypothèse 2*** : *les centres d'intérêt sont dépendants*
  Dans ce cas, la dépendance traduit l'existence d'un lien sémantique entre un sous ensemble de centres d'intérêt couvrant le même besoin en information que la requête. Ce lien sémantique peut s'exprimer à travers des relations de hiérarchisation des centres d'intérêt et permet ainsi le renforcement du degré de pertinence des documents. En accord avec le principe d'agrégation basée sur le consentement absolu (Fargier et Perny 2003), l'utilité globale d'un document sera obtenue par agrégation cumulative des valeurs des utilités élémentaires. L'opérateur d'agrégation proposé est en l'occurrence l'opérateur de la somme, défini comme suit :
  $$\Psi(\mu_1 ... \mu_n) = \sum(\mu_1 ... \mu_n) \tag{4.22}$$

4.5 Contribution au domaine de recherche

La spécificité de nos travaux ainsi que le cadre de déroulement de nos recherches dans ce domaine sont développés dans ce qui suit.

4.5.1 Positionnement de nos travaux vis-à-vis de la littérature

– *Evaluation de requêtes préférentielles :* à notre connaissance, nous sommes les précurseurs à l'utilisation du cadre formel offert par les graphes CP-Nets pour aborder la problématique du traitement des préférences conditionnelles de l'utilisateur lors de l'évaluation de la requête. Ces travaux ont abouti à deux résultats importants publiés dans des conférences internationales dans le domaine (Boubekeur et al. 2006; 2007; 2008), à savoir : (1) la définition d'un modèle d'appariement basé sur les graphes requêtes CP-Nets, (2) la définition d'un modèle sémantique d'appariement flexible requête préférentielle-document basés sur les graphes CP-Nets. On se démarque ainsi d'emblée des travaux connexes qui traduisent l'évaluation de ces requêtes par application de t-normes et co-normes à des expressions conjontives et/ou disjonctives. L'état de nos travaux, ne nous permet pas toutefois de mesurer précisément l'impact différencié de ces approches sur l'efficacité de la recherche ; c'est un travail que l'on inscrit dans nos perspectives de recherche.

– *Evaluation personnalisée de requêtes* : le résultat le plus important issu de nos activités de recherche dans ce cadre est sans doute la spécification d'un modèle formel d'accès personnalisé à l'information. Les premières propositions dans ce sens ont permis de définir formellement un modèle de RI comme un quadruplet $M =< I, U, E^t(U), P(I,U) >$ où I est le modèle de représentation de l'information (documents et requêtes), U est le modèle de représentation des centres d'intérêt de l'utilisateur, $E^t(U)$ est la fonction d'évolution de l'utilisateur en fonction du temps et $P(I,U)$ est la fonction d'évaluation de la pertinence d'une information relativement à un utilisateur décrit par ses centres d'intérêt, c'est précisément cette fonction qui est modélisée à travers le modèle inférentiel présenté dans ce chapitre. La première spécificité de nos travaux est d'exploiter le modèle du contexte (en l'occurrence les centres d'intérêt) dans la phase de l'appariement ; *à notre connaissance seul le modèle de (Lin et al. 2005b) a adopté ce principe avec, cependant, une approche fondamentalement différente* liée non pas à la spécification d'un problème décisionnel mais à la dérivation, de proche, en proche, de l'intention latente de l'utilisateur à travers les unités d'informations qu'il a manipulées.
Du point de vue de l'impact sur l'efficacité de la recherche, notre modèle a été comparé, selon un cadre d'évaluation approprié pour cette tâche (Cf. chapitre 5), à des modèles de base tels que *Okapi* et d'autres modèles personnalisés de référence dans le domaine tels que celui publié dans (Gauch et al. 2003). Les taux d'accroissement obtenus sont significatifs de l'ordre respectivement de 19% et 17%.

4.5.2 Support et structuration de nos travaux

Nos investigations de recherche autour de la problématique de l'évaluation contextuelle de la requête nous ont permis dans un premier temps, de cerner le foisonnement de verrous sous-jacents dans le cadre de l'*AS*

personnalisation de l'information[2] puis du projet *APMD* qui a été le support d'une grande partie de nos contributions dans le domaine. Nos travaux sur l'évaluation des requêtes préférentielles se sont articulés autour de la thèse de *Fatiha Boubekeur* que j'ai co-encadrée. Concernant nos travaux portant sur l'évaluation personnalisée de requêtes, les premières spécifications de ce modèle ont été publiées dans (Tamine et al. 2006a, Tamine et Boughanem 2006). De nouvelles spécifications y ont été apportées dans le cadre des travaux de master de *Amira Tifous* dans le but d'appliquer de nouveaux algorithmes de propagation de l'évidence dans le diagramme et définir de nouvelles mesures d'utilité. Ce travail a été poursuivi dans le cadre des travaux de thèse de *Nesrine Zemirli* que j'ai co-encadrée et qui a abouti à la mise en œuvre du prototype *SyriX* dont les résultats d'évaluation ont été publiés dans le journal international *Journal of Digital Information Management (JDIM)*(Tamine-Lechani et al. 2008) et conférences internationales (Tamine et al. 2007b;a). Ce prototype a également été l'un des résultats importants du projet APMD auquel nous avons participé.

4.6 Conclusion et perspectives

Dans ce chapitre, nous avons proposé des modèles et des stratégies d'évaluation de requête en contexte. Notre première contribution a porté sur l'utilisation des CP-Nets pour mettre en œuvre une évaluation flexible des requêtes. Les CP-Nets offrent un formalisme graphique simple et intuitif qui permet de structurer les préférences qualitatives de l'utilisateur de manière compacte puis de les évaluer automatiquement en utilisant des opérateurs d'agrégation.
Notre seconde contribution a porté sur la définition d'un modèle inférentiel pour l'évaluation de la requête tenant compte des centres d'intérêt de l'utilisateur qui l'a émise. Le modèle permet de supporter un processus de prise de décision quant à la présentation des documents et ce, en tenant compte de la mutuelle influence de nombreux critères : sujet de la requête, contenu des documents, portée des centres d'intérêt. Le modèle est spécifié selon une composante qualitative et composante quantitative traduisant les heuristiques connues en RI en l'occurrence : représentativité des termes dans un contenu d'information, critères de satisfaction d'une requête etc. L'algorithme de propagation de l'évidence mis en œuvre permet de relativiser de proche en proche l'estimation globale de la pertinence dans un contexte caractérisé par la multiplicité des centres d'intérêt.
L'ensemble de ces travaux, même si abordés sous deux angles différents, nous ont permis de mieux cerner les problèmes (et solutions) liés à l'intégration du contexte dans l'étape d'évaluation d'une requête.
Loin d'être achevés, nous envisageons de les poursuivre selon les perspectives suivantes :

- *Spécifier un modèle intégré qui exploite les deux facteurs du contexte dans le processus d'évaluation de la requête*. Notre objectif dans ce sens est de traiter aussi bien les préférences de l'utilisateur (en termes de

[2] AS 98 du CNRS 2003-2004

critères de recherche, de service,...) que ses centres d'intérêt dans le même algorithme d'évaluation de la pertinence. Nous considérons en effet, que les centres d'intérêt constituent une source d'évidence qui nous permet de mieux interpréter ses préférences et donc préciser la mesure de pertinence des résultats. Cette direction nous amènera à (1) spécifier un méta-modèle du contexte instanciable à l'aide de différents facteurs (2) définir des algorithmes de propagation d'évidence qui soient compatibles avec la spécificité de chaque facteur.

– *Améliorer la composante qualitative et composante quantitative de notre modèle d'évaluation personnalisée de requêtes*. En effet, en continuité avec nos travaux portant sur la modélisation du contexte cognitif (Cf. chapitre 2), nous envisageons d'intégrer dans notre modèle inférentiel d'évaluation de requête, des modèles cognitifs plus riches tels que des graphes de concepts plutôt que des structure plates de termes isolés. Sans remettre en cause la viabilité de notre modèle actuel, ce prolongement nous conduira à (1) définir de nouvelles distributions de probabilités permettant de traduire la représentativité des concepts dans les documents et centres d'intérêt (2) réviser nos opérateurs d'agrégation en passant par une meilleure caractérisation de la dépendance des centres d'intérêt.

– *Evaluer notre modèle flexible d'évaluation de requêtes préférentielles dans le cadre d'applications dédiées*. L'expression des préférences qualitatives dans les requêtes est particulièrement pratiquée dans des applications dédiées telles que le diagnostic médical et la recherche d'information dans le domaine juridique. Dans ce cadre, une requête est exprimée d'avantage sous forme d'une situation décrite par des conditions préférentielles. Le rôle du processus de recherche d'information est alors d'identifier l'information exprimant une situation analogue. Cette évaluation nous permettra de mesurer à la fois l'efficacité, l'efficience et la portée de nos contributions dans le domaine de l'évaluation flexible de requêtes préférentielles.

5 Evaluation des performances d'un système de RI contextuel

5.1 Introduction

L'objectif de l'évaluation d'un SRI est de mesurer ses performances. Les protocoles d'évaluation largement adoptés en RI sont empiriques et souvent basés sur une évaluation d'avantage quantitative que qualitative. Le modèle d'évaluation *Cranfield* (Cleverdon 1967a) est incontestablement le modèle de référence adopté dans les campagnes d'évaluation standards en RI. Ce modèle fournit une base d'évaluation comparative de l'efficacité de différents algorithmes, des techniques et/ou des systèmes moyennant des ressources communes : des collections de test contenant des documents, des requêtes préalablement construites et des jugements de pertinence associés construits selon la technique de *pooling*, des métriques d'évaluation essentiellement basées sur le rappel-précision. L'émergence de la RI orientée utilisateur a cependant remis en cause la viabilité de ce modèle pour l'évaluation de systèmes interactifs ou de manière générale, les systèmes d'accès contextuel à l'information (Ingwersen et Jarvelin 2005). Notre souci pour la validation de nos contributions dans le domaine nous a alors conduits à proposer, dans ce cadre, des méthodologies d'évaluation dont l'objectif est d'estimer l'efficacité des modèles d'accès contextuels à l'information que nous avons proposés. Le cadre d'évaluation proposé est basé sur l'enrichissement des ressources TREC (Voorhees 2001) par des contextes simulés. Plus précisément, un protocole issu de TREC *adhoc* est mis en œuvre pour évaluer notre modèle d'accès à l'information guidé par les centres d'intérêt de l'utilisateur (Cf. chapitre 4, section 4.4). Un autre protocole issu de *TREC-HARD* est mis en œuvre pour évaluer notre modèle exploitant des centres d'intérêt construits à partir des interactions utilisateur-SRI découpées en sessions de recherche (Cf. chapitre 2, section 2.3)

Ce chapitre est organisé autour de huit sections principales. La section 5.2 illustre la problématique de l'évaluation en RI contextuelle et par conséquent la nécessité de disposer d'une démarche d'évaluation rigoureuse et justifiable dans un tel cadre. La section 5.3. présente un état des lieux des méthodologies d'évaluation orientées contexte. La section 5.4 présente le cadre général, motivations et objectifs de notre approche d'évaluation. Les sections 5.5 et 5.6 sont dédiées respectivement à la description des

cadres d'évaluation *TREC adhoc* et *TREC-HARD*. La section 5.7 résume notre contribution dans ce domaine de recherche. Enfin, la section 5.8 conclut et présente les perspectives de nos travaux dans le domaine.

5.2 Problématique générale

L'évaluation classique des performances en RI est basée sur une approche de type laboratoire (*laboratory-based model*) initiée par *Cleverdon* (Cleverdon 1967a) dans le cadre du projet *Cranfield project II* ; c'est le cadre d'évaluation largement adopté dans les campagnes d'évaluation internationales telles que *TREC*, *INEX*[1] et *CLEF*[2]. Parmi les critères de performances, le rappel et la précision sont incontestablement les plus utilisés dans les travaux du domaine ; ils permettent d'évaluer la capacité du système à présenter des documents pertinents et à rejeter des documents non pertinents. Le paradigme de l'évaluation classique est fondé sur (1) l'utilisation d'une collection de test (2) l'estimation des performances du SRI grâce à des mesures d'évaluation (3) la comparaison de résultats obtenus entre algorithmes, modèles et techniques.
L'évaluation "à la Cranfield" a permis de faire avancer l'état de la recherche en RI, notamment la RI orientée-système, grâce aux nombreux avantages qu'elle offre, résumés dans ce qui suit :

- c'est un cadre d'évaluation qui garantit la reproductibilité des expérimentations et l'évaluation comparative des résultats qui en sont issus,
- le principe d'évaluation lié à la présence/absence des variables d'expérimentation, permet d'identifier les problèmes, de comprendre les origines et de cibler des solutions,
- les résultats issus des expérimentations sont généralisables dans le respect des conditions expérimentales,
- le cadre général permet d'asseoir des campagnes d'évaluation standards telles que *TREC* qui offrent diverses ressources partageables pour l'évaluation de nombreuses tâches en RI telles que le tâche de filtrage, la tâche question-réponse, la tâche de recherche sur le web etc.

Ce modèle d'évaluation classique, principalement issu de la RI orientée-système, a cependant été remis en cause par l'émergence de la RI orientée-utilisateur (Law et al. 2006, Kekalainen et Jarvelin 2004). On résume dans ce qui suit les limites majeures de ce type d'évaluation :

A. Les collections de test sont peu adéquates pour l'évaluation de la recherche d'information en contexte

1. *Les requêtes*
   Le protocole d'évaluation étant en mode *batch*, les requêtes sont supposées ainsi représenter à elles seules l'utilisateur. Les utilisateurs directs ayant émis ces requêtes, leurs centres d'intérêt et interactions

[1] INitiative for the Evaluation of XML retrieval
[2] Cross Language Evaluation Forum

avec le SRI ne font pas partie intégrante de la collection.

2. *Les jugements de pertinence*
   La pertinence considérée est thématique, indépendante du contexte, situation de recherche et centres d'intérêt des utilisateurs. Or, il a été bien montré (Borlund 2003a, Mizzaro 1998) que la notion de pertinence est plus complexe, couvrant des niveaux divers liés à la situation de recherche en cours : pertinence cognitive, pertinence affective, pertinence situationnelle, etc.

B. Les mesures d'évaluation ne sont pas exhaustives

1. *Portée*
   Les mesures de précision ne permettent pas d'évaluer des SRI opérationnels. En effet d'autres critères tels que le degré de couverture du besoin en information et la valeur ajoutée des résultats (cas de documents déjà vus par exemple) ne sont pas corrélées aux mesures de précision (Louise 1992). De plus, il n'a pas été prouvé que les différences statistiques entre le rappel et la précision de différents systèmes est significative dans des situations réelles de recherche d'information. A titre d'exemple, dans (Kim et Allen 2002, Law et al. 2006) des études expérimentales ont montré que les jugements des utilisateurs variaient de manière significative en fonction des SRI et des sujets de requête.

Ce constat critique a motivé les réflexions autour de méthodologies d'évaluation adaptées à une recherche d'information contextuelle. Dans la section suivante, nous en donnons un large aperçu.

5.3 L'évaluation de l'efficacité de modèles d'accès contextuel à l'information

Notons d'emblée, qu'il n'existe pas, à ce jour, de cadre standard pour l'évaluation de l'efficacité d'un modèle d'accès contextuel à l'information. Les travaux de validation en RI contextuelle reposent sur des méthodologies et cadres spécifiques en effectuant des choix (collections, jugements de pertinence, requête etc.) dépendant de la technologie mise en œuvre. L'analyse de ces travaux (Tamine-Lechani et al. 2009) fait émerger des approches d'évaluation synthétisées et discutées dans ce qui suit.

5.3.1 Approches d'évaluation

A. Les prémices dans TREC

L'évaluation de processus de recherche d'information en situation d'interaction a été initiée dans *TREC* à travers les tâches *Interactive et HARD*.

– *La tâche Interactive* (Harman 1995a). Cette tâche a été menée dans *TREC* depuis 1995 (TREC-4) jusqu'en 2002 (TREC-10). Elle a eu pour double objectif le développement de méthodologies appropriées à l'évaluation de processus de recherche d'information interactive et la mesure de l'impact des différentes caractéristiques des utilisateurs

dans l'évaluation de la pertinence des résultats. A cet effet, des questionnaires et des interviews sont établis au préalable pour décrire les utilisateurs participant effectivement à la campagne d'évaluation. Les principales caractéristiques recueillies concernent leur familiarité avec les sujets de la requête, leur expertise dans l'utilisation de moteurs de recherche, leur habilité à interagir avec le système, leur démarche pour mener une tâche de recherche d'information, etc. Les collections de documents n'ont pas subi de changements conséquents relativement aux campagnes *TREC* précédentes. Les mesures classiques de rappel-précision sont généralement utilisées pour mesurer l'effet de certaines variables isolées (liées aux caractéristiques citées ci-avant) sur les performances de recherche. A *TREC-7* une nouvelle mesure est proposée, en l'occurrence le rappel au niveau instance, qui mesure pour chaque utilisateur le nombre d'instances de réponses correctes trouvées pour une question donnée sur un intervalle de temps déterminé (15 à 20 minutes).

– *La tâche HARD* (Allan 2003a). Cette tâche a été menée dans *TREC* depuis 2003 (TREC-12) jusqu'en 2005 (*TREC-14*). Son objectif est d'atteindre les performances de haute précision pour des utilisateurs spécifiques. A cet effet, et à la différence de la tâche interactive, la tâche a introduit (1) l'utilisation de méta-données dans les documents et les requêtes de la collection tests qui décrivent le contexte de recherche, on cite notamment : la familiarité, le genre et la granularité (2) une pertinence graduelle (3) des mesures d'évaluation qui considèrent les niveaux de pertinence pour l'évaluation des performances de recherche ; d'autres informations additionnelles sont éventuellement demandées aux utilisateurs participant à la campagne d'évaluation à l'aide de formulaires de clarification. L'objectif de l'évaluation est alors de mesurer l'efficacité d'un système à fournir les réponses précises en fonction du contexte de la recherche.

Ces tâches *TREC* ont le mérite d'avoir introduit la dimension contexte dans le paradigme de l'évaluation. Cependant, les tâches sont très spécifiques, ne permettant pas d'évaluer, sans peine d'extensions, des modèles d'accès intégrant des dimensions du contexte allant au delà du *feedback* et des facteurs basiques du contexte tels ceux prévus dans la tâche *HARD*.

B. Approche basée sur la simulation du contexte

Cette approche est basée sur la définition de scénarios d'évaluation qui simulent des utilisateurs et interactions hypothétiques (Mostafa et al. 2003, Ryen et al. 2005). En clair, un scénario représente une situation de recherche qui met en œuvre un utilisateur simulé à travers ses centres d'intérêt (Sieg et al. 2007), ses interactions avec le SRI (Ryen et al. 2005) à travers des interfaces dédiées comme celles proposées dans *TREC* interactive. Généralement, des collections *TREC* (documents, requêtes et jugements de pertinence) et mesures dérivées du rappel-précision sont utilisées pour mesurer l'évaluation de l'efficacité du modèle. Cette approche présente deux avantages majeurs : (1) n'est pas couteuse en temps puisqu'elle n'im-

plique pas des utilisateurs réels (2) permet d'effectuer une évaluation comparative.

C. Approche basée sur des contextes réels

Cette approche implique la conduite d'expérimentations avec des utilisateurs réels, appelés participants, dans des situations de recherche typiques. L'objectif de l'évaluation est de mesurer l'efficacité de la recherche en tenant compte de la nature dynamique du besoin en information et du jugement de pertinence ainsi que de l'interaction utilisateur-SRI dans des situations de recherche réelles et bien spécifiées (*work task situation*) (Bystrom et Jarvelin 1995). Cette approche est de plus en plus adoptée dans le domaine (Liu et Yu 2004, Speretta et Gauch 2005, Lee et al. 2005a). Néanmoins, en plus du cout en temps induit, cette approche pose un problème lié à la reproductibilité des résultats et par conséquent à l'inconsistance de l'évaluation comparative. Plus précisément, il est difficile d'isoler les effets des utilisateurs eux même (centres d'intérêt, expertises, familiarité avec le sujet de la requête, etc.) de ceux du modèle en comparant les performances associées à différents scénarios d'évaluation. Dans le but d'y pallier, des recommandations sont proposées quant au déroulement des expérimentations et définition des scénarios d'évaluation (Chin 2001, Borlund 2003b).
Par ailleurs, en raison de l'absence de référentiel d'évaluation (pas de jugements de pertinences préalablement associés de manière intrinsèque aux documents résultant de l'évaluation d'une requête), les méthodologies d'évaluation inscrites dans le cadre de cette approche adoptent l'un ou l'autre des principes suivants :

1. utiliser les mesures classiques de rappel et de précision et leurs dérivées ; ceci est possible grâce à la construction d'un référentiel de pertinence sur la base de l'ensemble des documents jugés pertinents par les différents utilisateurs pour une même requête (Liu et Yu 2004, Ding et Patra 2007).
2. utiliser des mesures orientées rang telles que le rang moyen et *DCG* (*Discouted Cumulative Gain*) (Jarvelin et Kekalainen 2002) qui, indépendamment du rappel, se basent sur la position des documents pertinents dans la liste des résultats (Speretta et Gauch 2005, Agichtein et al. 2006).

5.3.2 Discussion

L'évaluation classique basée sur un modèle de laboratoire issue de la RI orientée-système et l'évaluation basée sur des contextes réels, issue de la RI orientée-utilisateur, délimitent le spectre des méthodes d'évaluation. Les deux approches présentent des avantages qui méritent d'être combinés, dans le cadre d'une évaluation dite *formative*, pour en accroître la robustesse (Petrelli 2008, Diaz et al. 2008). Dans ce sens, l'approche d'évaluation orientée-système peut en effet être recommandée dans le cadre des étapes préliminaires de conception d'un SRI dans le but d'évaluer les fonctionnalités de base : indexation, paramétrage, évaluation de la pertinence thématique etc. L'évaluation orientée-utilisateur est quant à elle re-

commandée lors des phases de "mise en service" des SRI impliquant des utilisateurs spécifiques avec un contexte cognitif (centres d'intérêt, préférences, connaissances etc.), des situations (but de la recherche, organisation sociale, lieu, temps etc..), des interactions. . . .
Entre ces deux approches, il existe une multitude de méthodologies d'évaluation développées dans la littérature de la RI contextuelle, proposant des collections de test, des protocoles et des mesures d'évaluation. On déplore à ce jour, l'absence de protocoles standards permettant d'être réutilisés à des contextes d'utilisation différents. Ceci reste un réel défi et une piste de recherche que nous inscrivons dans nos perspectives.
Confrontés à ce défi dans le cadre de nos activités de validation, nous avons alors proposé des méthodologies d'évaluation pour étalonner l'efficacité des modèles d'accès à l'information que nous avons proposés.

A cet effet, nous avons adopté une approche basée sur des contextes simulés considérés comme des ressources additionnelles au cadre d'évaluation *TREC*. Notre choix est justifié par (1) la faisabilité de la validation à moindre cout (2) notre retour d'expérience des campagnes d'évaluation *TREC*.
Dans les sections suivantes, nous spécifions les cadres d'évaluation que nous avons développés et effectivement exploités pour valider nos travaux.

5.4 Notre approche d'évaluation : cadre général, motivations et objectifs

La problématique générale de l'évaluation de l'accès contextuel à l'information, résumée en section 5.2., se décline dans le cadre de nos activités, à travers les points suivants :

1. Il n'existe pas de cadre d'évaluation standard adapté à un accès contextuel à l'information, capable de montrer la viabilité de nos contributions quant à l'intégration des centres d'intérêt dans le modèle de recherche d'information, la définition des sessions de recherche pour l'apprentissage des centres d'intérêt, la qualité du réordonnancement déployé dans le but d'améliorer la précision de la recherche,

2. les tâches *TREC*, en l'occurrence Interactive et *HARD*, ne nous permettent pas à moins d'une extension couteuse en temps de mise en œuvre, de construire de manière explicite une composante centres d'intérêt, de délimiter des contextes d'usage liés à des sessions de recherche,

3. il est difficile d'étalonner la pertinence des résultats issus des modèles que nous avons proposés, sans cibler des domaines d'intérêt spécifiques avec des jugements de pertinence qui sont issus, non pas d'une requête ponctuelle, mais d'un historique de requêtes,

4. même si l'évaluation basée sur des situations de recherche réelles impliquant des participants est sans doute la plus significative (en

garantissant le respect des recommandations pour le déroulement de telles expérimentations), elle nous semblait peu envisageable en raison de son cout de mise en œuvre, en l'état de nos recherches.

Pour l'ensemble de ces raisons, une première solution envisagée était de proposer des méthodologies d'évaluation fondées sur une approche de contextes simulés, plutôt que des contextes réels. Notre expérience, acquise à travers notre participation aux campagnes d'évaluation orientée-laboratoire (Boughanem et al. 2002b, Pinel-Sauvagnat et Boughanem 2004, Baziz et al. 2005b), nous ont alors motivés et confortés dans l'idée de réutiliser particulièrement, le cadre *TREC* pour atteindre nos principaux objectifs, résumés dans ce qui suit :

- évaluer l'efficacité du modèle d'accès contextuel à l'information guidé par le contexte cognitif de l'utilisateur, plus précisément ses centres d'intérêt,
- déterminer l'impact de la construction de ces centres d'intérêt à partir de sessions de recherche corrélées par le sujet de la requête,
- situer l'apport de notre modèle par comparaison à d'autres modèles d'accès contextuel de référence (comparables) dans le domaine.

Ces objectifs nous ont guidés dans la définition de deux cadres d'évaluation issus de *TREC*, proposés en séquence, de manière corrélée à l'avancement de nos travaux. Ces cadres, en l'occurrence *TREC ad-hoc* et *TREC HARD* seront spécifiés dans les sections suivantes.

5.5 Le cadre d'évaluation issu de TREC *ad-hoc*

Ce cadre d'évaluation a été initialement défini pour l'évaluation de l'accès contextuel guidé par le profil cognitif de l'utilisateur, basé mots clés (Cf. chapitre 2, paragraphe 2.3.1) ; il a été étendu pour supporter des profils cognitifs basés sur des graphes de concepts (Cf. chapitre 2, paragraphe 2.3.2). Indépendamment de la technologie à évaluer, le cadre d'évaluation requiert :

1. Une collection de test composée :
   - de requêtes dont on connaît *a priori* (1) le sujet général, (2) un ensemble de documents pertinents. Ces requêtes sont à la base de l'apprentissage des centres d'intérêt d'utilisateurs hypothétiques,
   - d'une collection de documents interrogée par ces requêtes.
2. un algorithme de simulation des centres d'intérêt des utilisateurs selon le principe suivant : dériver pour chaque sujet, considéré comme un centre d'intérêt possible d'un utilisateur, une description de son contenu. Cet algorithme manipule en entrée (1) la description de la requête qui couvre le sujet en cours, (2) les descripteurs des documents pertinents associés,
3. un mécanisme de validation croisée est mis en place afin de générer

plusieurs descriptions candidates pour un centre d'intérêt donné, évitant ainsi le biais de l'évaluation.

4. des métriques orientées rappel-précision

Nous avons spécifié et mis en œuvre ce cadre d'évaluation en utilisant particulièrement une collection *TREC-Adhoc* comme décrit ci-après.

5.5.1 Collection de test

A. Requêtes

Nous avons utilisé les requêtes de la collection *TREC 1* numérotées de 51 à 150 ($q_{51} - q_{150}$). Le choix de cette collection de requêtes est guidé par le fait qu'elles sont annotées d'un champ particulier noté « **Domain**» qui décrit un domaine d'intérêt traité par la requête. C'est à juste titre, cette méta-donnée qui sera exploitée pour simuler des utilisateurs hypothétiques avec des centres d'intérêt issus de ces domaines. Ces requêtes couvrent huit (8) domaines illustrés sur la figure 5.1 et exploités pour nos différentes expérimentations. Le format d'une requête est le suivant :

```
<top>
<head> Tipster Topic Description
<num> Number: 062
<dom> Domain:  Military
<title> Topic:  Military Coups D'etat
<desc> Description: Document will report a military
coup d'etat, either attempted or successful,
in any country.
<smry> Summary: Document will report a military
coup d'etat, either attempted or successful, in any country.
</top>
```

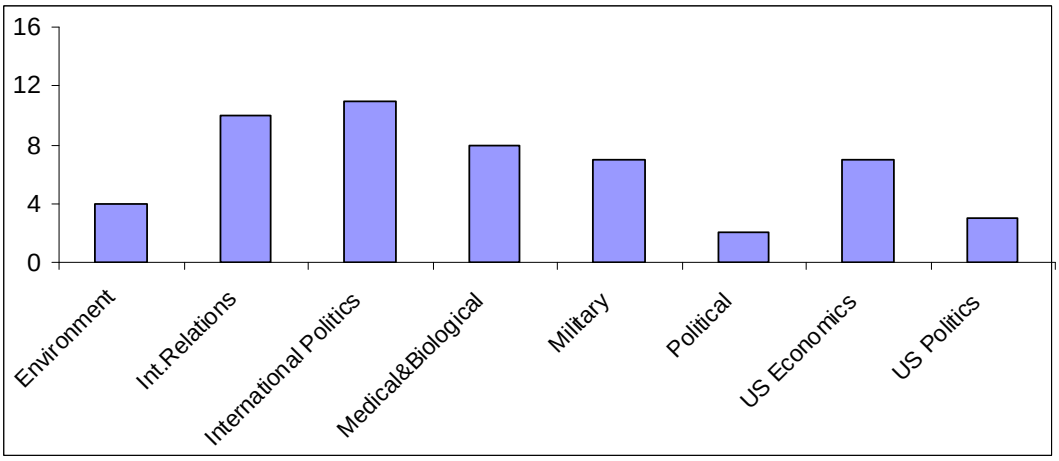


FIG. 5.1 – *Distribution des requêtes par domaine d'intérêt*

Notons que les variations des nombres de requêtes par domaine, nous permet d'évaluer notre modèle d'accès, dans différentes conditions liées à la "longueur" de l'historique de requêtes simulés pour une même session.

TAB. 5.1 – *Caractéristiques de la collection de test TREC 1*

Nombre de domaines	8
Nombre de documents	3576208
Nombre de requêtes	52
Nombre de termes distincts	589212
Longueur moyenne d'un document	53,64
Longueur moyenne d'une requête	3,5

On pourra ainsi se rapprocher de conditions réalistes où le centre d'intérêt d'un utilisateur est appris à partir de peu ou prou de requêtes.

B. Collection de documents

La collection de test de la campagne d'évaluation *TREC 1 ad-hoc* utilisée dans ce cadre, est celle des disques 1, 2 et 3, qui est à juste titre, interrogée par les requêtes ($q_{51} - q_{150}$) décrites précédemment. Les documents de cette collection sont issus de différents articles de presse tels que *Associate Press (AP)*, *Wall street journal (WJS)*, *Financial times*. Les caractéristiques de la collection sont présentées dans la tableau 5.1.

C. Les centres d'intérêt

Les centres d'intérêt représentent un élément initialement absent de la collection *TREC*. Cette ressource additionnelle a été intégrée dans la collection de test grâce un algorithme de simulation qui les génère à partir des documents pertinents (jugés pertinents par les assesseurs) associées à des requêtes de même domaine. Plus précisément, les étapes de ce processus de simulation sont les suivantes :

1. Pour chaque domaine k de la collection (noté Dom^k avec $k = (1..8)$), nous sélectionnons parmi les n requêtes associées à ce domaine, noté Q^k, un sous-ensemble de $n - 1$ requêtes pour constituer l'ensemble d'apprentissage du centre d'intérêt, noté C_n^k.
2. A partir de cet ensemble d'apprentissage, on extrait automatiquement la liste des vecteurs[3] documents pertinents et non pertinents associés à chaque requête.
3. Partant de ces vecteurs documents, le processus de construction des centres d'intérêt est déployé. Du point de vue technologique, ces vecteurs sont à la base de l'apprentissage de centres d'intérêt :
   – basés mots-clés et ce grâce à un algorithme qui permet d'agréger l'information pertinente tel que l'algorithme d'*OKAPI* (Robertson et al. 1995),
   – ou alors basés sur des graphes de concepts issus d'une ontologie de référence
4. On réitère le processus n fois pour chacun des n centres d'intérêt associés au domaine k. A chaque itération, on fait varier le sous ensemble d'apprentissage pris parmi n requêtes associées au domaine k en sélectionnant une nouvelle requête à chaque fois.

[3]Le vecteur document est composé de poids associés à chacun des termes d'indexation

5.5.2 Stratégie de test

La simulation d'utilisateurs hypothétiques représentés par de centres d'intérêt générés selon le processus décrit précédemment, est subordonnée, de surcroît, à l'application d'une stratégie de test basée sur la *validation croisée*. Cette dernière, consiste globalement à faire varier l'ensemble des requêtes d'apprentissage servant de base à la construction d'un centre d'intérêt, en l'alternant itérativement avec l'ensemble des requêtes à tester. L'intérêt de cette stratégie de test est double :

1. simuler la variation des centres d'intérêt pour un utilisateur hypothétique donné, par génération et application de différents centres d'intérêt issus d'un même domaine testé. Ceci permet de traduire, sous l'angle d'une démarche basée sur les contextes simulés, l'évolution dynamique des centres d'intérêt dans des situations de recherche différentes,

2. évaluer un ordre de grandeur des performances de recherche plutôt que des performances obtenues dans une situation spécifique de recherche. Cette évaluation est obtenue en faisant la moyenne des performances obtenues sur les différents tests effectués.

La stratégie de test par validation croisée est illustrée par l'algorithme suivant :

Algorithme 3 Test par validation croisée

Entrée/Sortie : Requêtes test/apprentissage
Pour chaque requête Q^k associée au domaine C_k **Faire**
-Subdiviser l'ensemble des n requêtes du domaine en $q_{apprentissage}$ en sous-ensemble d'apprentissage de $n-1$ requêtes
et en q_{test} sous-ensemble de test contenant la n^{ime} requête
-Apprendre les centres d'intérêt à partir de $q_{apprentissage}$ soit C
-Tester les requêtes q_{test} avec C
Fin pour

5.5.3 Métriques d'évaluation

Le cadre d'évaluation étant issu de *TREC*, nous avons utilisé les mesures classique de la précision pour les X premiers documents restitués (PX) et de la MAP (Mean Average Precision) la moyenne de toute les précisions. Ces deux métriques permettent d'évaluer les performances de recherche selon deux aspects :

- PX : est la proportion de documents pertinents dans les X premiers documents retrouvés. Elle permet d'exprimer la satisfaction de l'utilisateur vis-à-vis des X premiers résultats pertinents. Elle constitue ainsi une mesure importante pour l'évaluation de la haute précision. Dans notre cas, on retient les précisions pour les 5 et 10 premiers documents : $P5$ et $P10$.
- MAP : est la précision moyenne pour l'ensemble des documents pertinents retournés. Elle exprime la capacité du modèle à sélectionner

les documents pertinents en réponse à toutes les requêtes expérimentées.

5.5.4 Mise en œuvre du cadre d'évaluation *TREC-adhoc*

Nous avons mis en œuvre le cadre d'évaluation *TREC-adhoc* et l'avons exploité particulièrement pour valider notre modèle d'accès à l'information guidé par les centres d'intérêt de l'utilisateur. Les principaux résultats obtenus portent notamment sur (1) l'évaluation l'efficacité de notre modèle comparativement à des modèles de base non guidés par les centres d'intérêt (2) l'évaluation de l'efficacité de notre modèle comparativement à d'autres modèles d'accès guidés par le contexte.

A. Evaluation comparative avec le modèle *OKAPI*

Une première initiative de validation consistait à mesurer la capacité de notre modèle à exploiter efficacement l'évidence issue des centres d'intérêt de l'utilisateur pour sélectionner l'information. Une évaluation comparative avec le modèle de référence *OKAPI* a montré que le modèle atteignait des taux d'accroissement[4] significatifs des performances moyennes de $19,51\%$ et $27,30\%$ respectivement à $P10$ et MAP ; les taux sont variables cependant en fonction du centre d'intérêt couvert. Le tableau 5.2 présente les résultats obtenus.

	Modèle Okapi		**Notre modèle**	
Domaine	*P10*	*MAP*	*P10*	*MAP*
Environment	0,47	0,29	$0,62^{\triangle}$	$0,29^{\diamond}$
Inter. Relations	0,21	0,06	$0,32^{\diamond}$	$0,12^{\triangle}$
Inter. Politics	0,20	0,08	$0,24^{\triangle}$	$0,12^{\triangle}$
Medical-Biological	0,38	0,06	$0,35^{\triangledown}$	$0,11^{\triangle}$
Military	0,27	0,10	$0,37^{\triangle}$	$0,15^{\triangle}$
Political	0,05	0,00	$0,20^{\triangle}$	$0,01^{\triangle}$
US Economics	0,28	0,10	$0,34^{\triangle}$	$0,12^{\triangle}$
US Politics	0,5	0,06	$0,40^{\triangledown}$	$0,07^{\triangle}$
Moyenne	0,29	0,09	$0,35^{\triangle}$	$0,12^{\triangle}$
Accroissement	**+19,51%**	**+27,30%**		

TAB. 5.2 – *Evaluation comparative avec le modèle Okapi*

En outre, une analyse plus fine des résultats focalisée sur les requêtes difficiles, dont les précisions de recherche $P10$ ne dépassent pas $0,3\%$, montre que notre modèle est à l'origine d'un accroissement des performances, comme illustrée sur le tableau 5.3.

B. Evaluation comparative avec le modèle de Gauch (Speretta et Gauch 2005)

Dans une seconde étape, nous avons procédé à l'évaluation comparative de notre modèle avec le modèle de *Gauch* (Speretta et Gauch 2005) dans le

[4]Le symbole $\triangle$ signifie que l'accroissement est positif, $\triangledown$ signifie que l'accroissement est négatif, $\diamond$ que l'accroissement n'est pas significatif

P@10 à la baseline	Nombre de requêtes	Accroissement
$[0\ \ 0,1[$	11	+3,63%
$[0,1\ \ 0,2[$	10	+110%
$[0,2\ \ 0,3]$	7	+11,11%
	28	$41,58\%^{\triangle}$

TAB. 5.3 – *Evaluation des performances des requêtes difficiles*

but de nous positionner relativement à d'autres modèles de l'état de l'art basés sur l'exploitation des centres d'intérêt de l'utilisateur pour estimer la pertinence de l'information. Ce modèle combine linéairement les rangs des documents obtenus par appariement avec la requête d'une part et le profil d'autre part. Ce dernier est composé de l'ensemble des concepts les mieux pondérés, suite à la projection des documents jugés pertinents sur l'ontologie. Cette évaluation a été possible en utilisant le cadre d'évaluation que nous avons défini, se basant toutefois sur une représentation des centres d'intérêt basés mots-clés pour rendre nos résultats comparables. Le tableau 5.4 résume les résultats obtenus.

	Modèle de Gauch		**Notre modèle**	
Domain	*P@10*	*MAP*	*P@10*	*MAP*
Environment	0,37	0,24	$0,62^{\triangle}$	$0,29^{\triangle}$
Inter. Relations	0,25	0,1	$0,32^{\triangle}$	$0,12^{\triangle}$
Inter. Politics	0,34	0,13	$0,24^{\triangledown}$	$0,12^{\triangledown}$
Medical-Biological	0,21	0,06	$0,35^{\triangle}$	$0,11^{\triangle}$
Military	0,38	0,13	$0,37^{\diamond}$	$0,15^{\triangle}$
Political	0,05	0,01	$0,20^{\triangle}$	$0,01^{\diamond}$
US Economics	0,27	0,14	$0,34^{\triangle}$	$0,12^{\triangledown}$
US Politics	0,53	0,11	$0,34^{\triangledown}$	$0,12^{\triangledown}$
Moyenne	0,30	0,11	$0,35^{\triangle}$	$0,12^{\triangle}$
Taux d'accroissement			**+17,89%**	**+7,6%**

TAB. 5.4 – *Evaluation comparative avec le modèle de Gauch*

5.6 Le cadre d'évaluation issu de TREC *HARD*

Du point de vue chronologique, ce cadre d'évaluation a succédé au cadre d'évaluation *TREC ad-hoc*. Il a eu pour point de départ deux principaux enseignements issus du premier cadre d'évaluation qui ont constitué également des motivations pour la définition de ce second cadre d'évaluation. Nous les résumons dans ce qui suit :

1. la démarche d'évaluation basée sur les utilisateurs hypothétiques augmentant une collection *TREC* est viable et rigoureuse du point de vue des normes d'évaluation. Néanmoins, est elle généralisable à d'autres types de requêtes, en l'occurrence difficiles ?

2. le principe de construction des centres d'intérêt avec des domaines bien identifiés, nous permet d'estimer les performances de notre modèle pour des situations de recherche traitant à la fois d'un

unique sujet générique. Qu'en est il de la précision de la recherche par variation des sujets de requête non clairement annotées d'un domaine d'intérêt (scénario plus proche de la réalité) ?

Le premier point a guidé notre choix pour l'utilisation de la collection *TREC HARD*, cette collection contient des requêtes particulièrement difficiles (Allan 2003b) ; le second nous a conduits à définir des protocoles nous permettant de simuler des sessions de recherche puis d'évaluer le modèle d'accès contextuel à l'information dans des situations de recherche impliquant : (1) des utilisateurs hypothétiques représentés par des centres d'intérêts simulés à partir d'un historique de requêtes, (2) des sessions de recherche identifiées par basculement du sujet général.

Ce cadre d'évaluation a été spécifié et mis en œuvre en s'articulant sur une collection *HARD* (comportant des requêtes particulièrement difficiles), cependant les stratégies de construction des centres d'intérêt et de test proposées sont généralisables pour toute collection de test comportant des requêtes et des jugements de pertinence associés. Nous présentons dans ce qui suit les principales caractéristiques de ce protocole d'évaluation.

5.6.1 Collection de test

A. Requêtes

Les requêtes *HARD* respectent le format général des requêtes (*topics*) mais comportent en plus, des métadonnées qui permettent de décrire le contexte de recherche, comme illustré dans ce qui suit dans un extrait de la requête *HARD-233* :

```
<num> HARD-233</num>
<title> Human rights</title>
<descr>
In what ways do international advocacy groups
demonstrate their concern over human rights in China?
</descr>
<narr>
Examples of on-topic documents may include the following:
1. Amnesty International (human rights groups)
2. Human rights in China(US based group)
3. Comments of Chris Patten
4. China's role in Tibet
5. political dissidents
6. Summary executions, torture and arbitrary arrests
7. Women's rights groups
</narr>
<searchterms>
Amnesty International, political dissidents, women rights
groups, China, human rights
</searchterms>
```

```
<hard> item=purpose, value=Background</hard>
<hard> item=genre, value=Overview</hard>
<hard> item=granularity, value=Document</hard>
<hard> item=familiarity, value=2</hard>
<relt>
```

Cependant, le cadre d'évaluation que nous définissons n'exploite pas pas ces métadonnées pour construire les contextes de recherche. Ce dernier est défini à travers (1) les centres d'intérêt construits à partir des sujets des requêtes (2) des sessions de recherche simulées selon un séquencement approprié de requêtes définissant des basculements dans les sujets de requête. Ainsi, la notion de sujet de requête est la base fondamentale pour la construction du contexte. Dans le but, de simuler un historique de requêtes traitant du même sujet, nous avons fait le choix, de subdiviser chaque requête, traitant *a priori* d'un sujet principal, en sous-requêtes traitant de sujets connexes. Ces deux éléments du contexte, développés dans les paragraphes suivants, sont conçus à la base, par définition de sous-requêtes issues d'une requête *HARD*. Le processus de constitution de la collection de requêtes test est basé sur l'extraction de 3 sous-requêtes à partir de chacune des 30 requêtes *HARD*, qu'on qualifiera de principales, selon les étapes suivantes :

1. Extraire le profil *pertinence* de la requête principale q en construisant l'ensemble des N vecteurs documents pertinents associés, soit dp_q,
2. Subdiviser ce profil en p sous-profils, notés sp_i, $sp_i \subset dp_q$,
3. Pour chaque sous-profil *pertinence* sp_i, créer un vecteur centroïde selon la formule : $c_i(t) = \frac{1}{|sp_i|} \sum_{d \in sp_i} w_{td}$, w_{td} est le poids du terme t dans le document d,
4. Extraire de chaque centroïde la sous-requête représentée par les k termes les mieux pondérés,
5. Eliminer les documents pertinents dp_q de la requête de la collection de test.

Dans le but de montrer la fiabilité de notre processus d'extraction de sous-requêtes, nous avons évalué :

- le taux de recouvrement de chaque sous-requête relativement à la requête principale. Ce taux est calculé par estimation du pourcentage de documents pertinents communs retournés par ces deux types de requêtes. La figure 5.2 montre bien que les sous-requêtes permettent de retourner autant, sinon plus de documents pertinents que la requête principale, ce qui traduit bien que les sous-requêtes traitent du sujet de la requête principale,

- le taux de non recouvrement de chaque sous-requête relativement à la requête principale. Ce taux est calculé par estimation du pourcentage de documents différents retournés par chaque type de requêtes et classé parmi les 20 ou 50 premiers documents retournés. La figure 5.3 montre bien , avec un taux de recouvrement de plus de 40% que

TAB. 5.5 – *Caractéristiques de la collection de test HARD*

Nombre de documents	1033461
Nombre de requêtes	30
Nombre de termes distincts	592373
Longueur moyenne d'un document	219.894952

les sous-requêtes traitent de sujets différents, ce qui va dans le sens de la complétude du sujet traité par la requête principale.

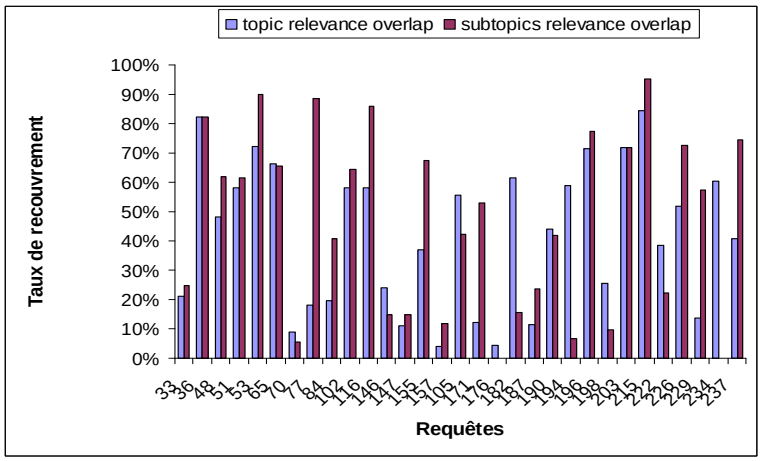


FIG. 5.2 – *Taux de couverture des requête*

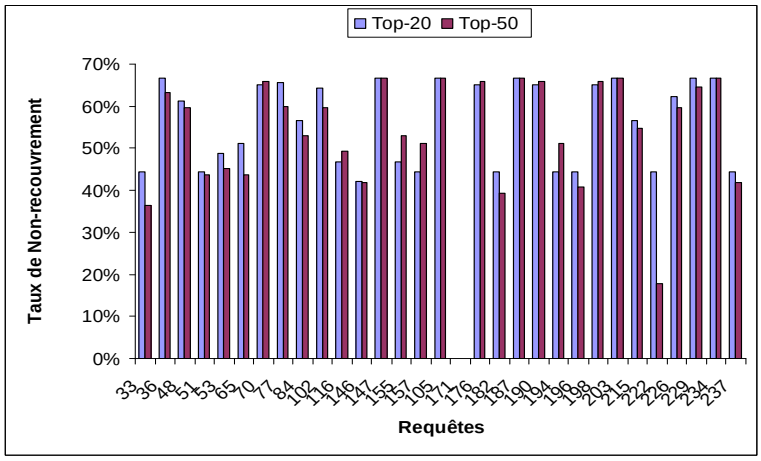


FIG. 5.3 – *Taux de non couverture des requêtes*

B. Collection de documents

Le corpus HARD comprend des documents comprenant des textes issus du *NewsWire 1999, AQUAINT corpus et U.S. government* . Le tableau 5.5 décrit les caractéristiques de cette collection.

C. Les centres d'intérêt

Le principe de construction des centres d'intérêt est analogue à celui décrit en 5.5.1. Dans ce cadre précisément, (1) la notion de domaine, clairement

identifié dans le cas de la collection *TREC ad-hoc* est remplacée par la notion de sujet de requête principal, non connu *a priori*, (2) les requêtes associées aux domaines, sont remplacées par les sous-requêtes associées à la requête principale en cours de traitement.

D. Les sessions de recherche

Rappelons qu'une session de recherche est constituée d'un ensemble de requêtes traitant d'un même sujet dominant. Dans le cas de ce protocole, une session est représentée par l'ensemble des sous-requêtes associées à une requête principale. Les sessions de recherche sont également simulées selon un processus guidé par le basculement dans les sujets de requêtes. Comme les situations de recherche sont hypothétiques, notre volonté consistait à définir un scénario qui soit le plus générique possible pour donner un ordre de grandeur plausible des performances proches de la réalité. A cet effet, nous avons défini un séquencement dit critique, créé à partir de l'ensemble des requêtes test principales Q, selon le processus suivant :

1. initialiser le séquencement critique $S = \oslash$,
2. choisir aléatoirement une requête principale $q_i \in Q$ puis aligner ses sous-requêtes q_i^1, q_i^2, q_i^3dans le séquencement S, $S = S \cup \{q_i^1,\ q_i^2,\ q_i^3\}$,
3. calculer la corrélation thématique (Cf chapitre 2) de la requête q_i à chaque requête $q_j \in Q\ \ j \neq i$, soit $Corr(q_i, q_j)$,
4. retenir $q_j^* = argmax(Corr(q_i, q_j))$,
5. aligner les sous-requêtes associées à q_j^* dans S, $S = S \cup \left\{q_j^{*1},\ q_j^{*2},\ q_j^{*3}\right\}$,
6. $Q = Q - q_j^*$, $q_i = q_j^*$,
7. si $Q \neq \oslash$ aller à 3.

L'alignement des sessions de recherche fondé sur la corrélation thématique maximale entre requêtes successives est motivée par notre volonté de confronter nos évaluations expérimentales à des sessions de recherche caractérisées par un basculement de sujet général éventuellement difficile à identifier.

E. Stratégie de test

La stratégie de test est basée sur la séparation entre ensemble de requêtes d'apprentissage et ensemble de requêtes test. L'ensemble des requêtes d'apprentissage permet (1) de créer les centres d'intérêt, (2) de définir les seuils de corrélation permettant de créer les sessions de recherche. L'ensemble de test comprend les requêtes effectivement testées pour l'évaluation des performances du modèle. De surcroît, notons que les documents pertinents ayant servi à créer les centres d'intérêt lors de la phase d'apprentissage ne sont pas considérés pour l'évaluation des performances associées aux requêtes test. Ceci permet en effet de ne pas biaiser les résultats dans le sens des documents pertinents déjà considérés dans le protocole d'évaluation.

F. Métriques d'évaluation

De même que pour le protocole précédent, nous utilisons des métriques classiques dérivées du rappel-précision. Nous avons particulièrement utilisé la Top_n *Précision* et le Top_n *Rappel* calculés comme suit :

$$Top - nPrecision = \frac{RelDoc_n}{n} \tag{5.1}$$

$$Top - nRappel = \frac{RelDoc_n}{RelDoc_{total}} \tag{5.2}$$

où $RelDoc_n$ est le nombre de documents pertinents apparaissant parmi les n premiers, $RelDoc_{total}$ est le nombre total de documents pertinents associés à la sous-requête (après élimination du profil requête ayant servi à sa création)

5.6.2 Mise en œuvre du cadre d'évaluation *TREC-HARD*

Ce cadre d'évaluation a été mis en œuvre puis exploité pour montrer globalement la viabilité des contextes cognitifs créées à partir d'un historique de recherche vu comme un ensemble de sessions de recherche ciblant des sujets généraux. Plus précisément, nous avons procédé en deux grandes étapes :

1. le cadre d'évaluation est tout d'abord exploité pour mesurer la précision de la classification des sessions de recherche selon la corrélation thématique des sujets abordés,
2. le cadre d'évaluation a été ensuite exploité pour mesurer l'efficacité du modèle d'accès basé sur le contexte cognitif ainsi défini et ce par comparaison au modèle de base *OKAPI* puis au modèle de *Gauch* (Speretta et Gauch 2005).

A. Evaluation de la classification des sessions de recherche

Dans le but d'atteindre cet objectif, nous avons calculé sur un historique de recherche simulé (Cf. section 5.6.1), calculé des corrélations thématiques entre requêtes successives puis fixé en fonction des mesures utilisées (Cf. chapitre 2, paragraphe 2.3.2) les meilleures valeurs de seuils, en ce sens qu'elles nous permettent d'atteindre les meilleurs taux de classification. Les figures 5.4 et 5.5 nous montrent ces taux sur la base de l'utilisation des mesures de *WebJaccard* et mesure de *Kendall*.

Les résultats montrent que notre protocole nous permet effectivement d'identifier des jalons des sessions avec des taux de précision significatifs pour la mesure de *Kendall*. C'est un premier résultat qui n'a été possible que grâce à la spécification de la "session de recherche" comme ressource additionnelle dans la collection de test.

B. Evaluation de l'efficacité du modèle d'accès

L'évaluation de l'efficacité d'un modèle classique basé sur le réordonnancement des résultats de recherche utilisant le contexte cognitif de l'utilisateur s'est effectué par comparaison (1) à une *baseline OKAPI* n'exploitant

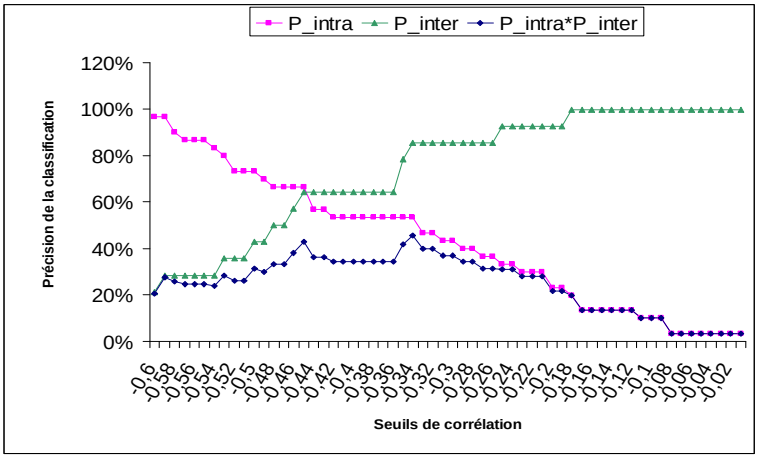


FIG. 5.4 – *Précision de la classification basée sur la mesure de Kendall*

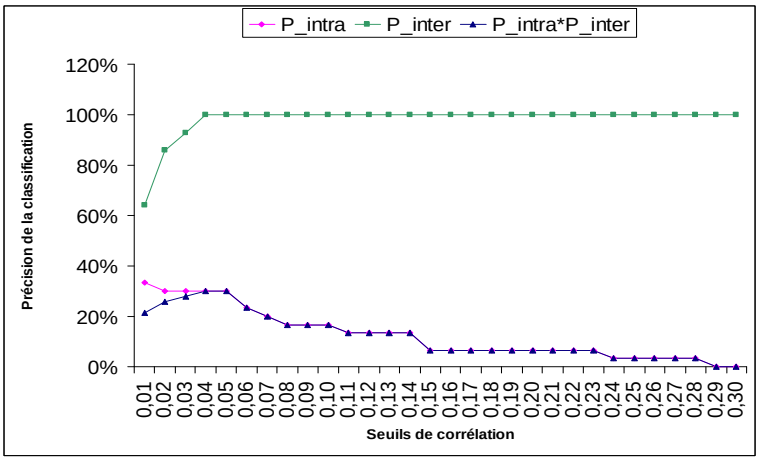


FIG. 5.5 – *Précision de la classification basée sur la mesure de WebJaccard*

pas le contexte (2) au modèle de *Gauch* exploitant cette fois (relativement à la comparaison effectuée selon le cadre d'évaluation *TREC adhoc*) des contextes cognitifs sémantiques comme spécifié dans l'approche publiée dans (Speretta et Gauch 2005). Les figures 5.6 et 5.7 montrent un avantage significatif pour notre modèle aussi bien selon la mesure du rappel que de la précision sur les n premiers documents présentés à l'utilisateur.

5.7 Contribution à ce domaine de recherche

La définition de cadres de validation empirique pour la validation de modèles et stratégies de RI contextuelle constitue un résultat important en soi, d'autant plus important qu'il a permis de faire asseoir voire faire avancer l'ensemble de nos contributions dans le domaine, présentées dans les chapitres précédents. Dans ce qui suit, nous mettons en exergue l'aspect novateur de nos travaux puis précisons le support qui nous a permis de les mener.

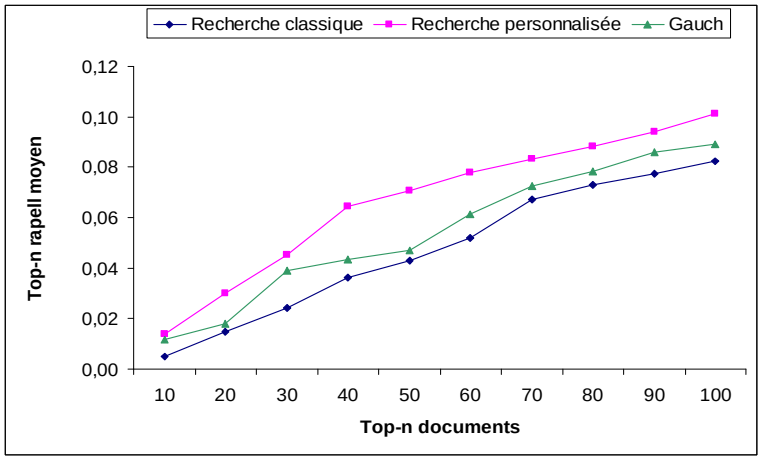


FIG. 5.6 – *Evaluation comparative selon la mesure du Top-n rappel*

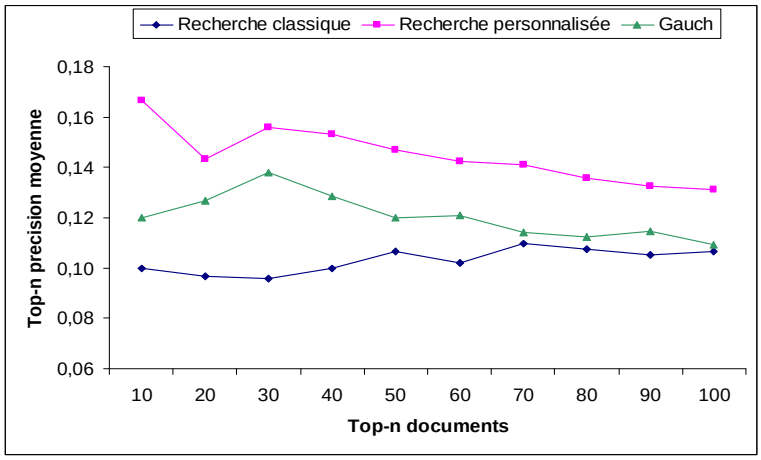


FIG. 5.7 – *Evaluation comparative selon la mesure du Top-n précision*

5.7.1 Positionnement de nos travaux vis-à-vis de la littérature

La littérature relative à l'évaluation contextuelle fait clairement état de l'absence de cadre d'évaluation partageable. Les cadres ayant servi à la validation des travaux du domaine sont généralement basés sur des contextes simulés à partir d'interfaces dédiées ou de campagnes menées avec des utilisateurs réels, au travers d'expérimentations non reproductibles. Hormis les tâches spécifiques *Interactive et HARD*, nous sommes parmi les précurseurs (avec (Bai et al. 2007)) à proposer une approche de validation basée sur des contextes simulés issus d'une collection *TREC*. Nos objectifs étant toutefois différents (dérivation de descriptions statiques de domaines pour la clarification de la requête dans le cas des travaux de (Bai et al. 2007)) nous sommes, à notre connaissance, les seuls à avoir proposé une méthodologie de simulation des centres d'intérêt de l'utilisateur par découpage thématique des sessions de recherche en utilisant des ressources *TREC*.
Les cadres que nous avons définis ont le mérite d'être reproductibles, à moindre coût de mise en œuvre, donnant un ordre de grandeur fiable des performances des modèles. Loin d'être idéaux, ces cadres méritent d'être

cependant étendus, voire révisés pour de nombreuses raisons que nous développerons en perspectives.

5.7.2 Structuration et support de nos travaux

Ces résultats importants, qui ont fait l'objet de nombreuses publications d'audience internationale et nationale dont (Tamine-Lechani et al. 2009, Tamine et al. 2008a, Daoud et al. 2008a;c), constituent l'aboutissement de nombreuses initiatives de recherche.
La première concerne notre participation active au projet *APMD* qui a constitué un cadre à travers lequel nous avons davantage pris conscience des spécificités de la problématique de l'évaluation dans le cas de la RI contextuelle, vue sous des angles très divers émanant de deux communautés différentes BD et RI : spécification des besoins en termes de mesures de performances, définition de méthodologies d'évaluation capables de considérer une ou plusieurs dimensions du contexte, extraction de contextes à partir de volumes de données etc. Ces travaux ont été sanctionnés par la concetion et mise en œuvre de la plate-forme *APMD-WORKBENCH* qui regroupe des jeux de données dérivés du cadre *TREC adhoc* et le second des bases *IMDb et MovieLens*.

Ce cadre d'évaluation a également servi à la validation des travaux de thèse de *Nesrine Zemirli* que j'ai co-encadrée qui ont abouti à terme à développer le prototype *SiRiX* comme extension au noyau *Mercure*. Ce cadre a été en outre exploité pour la validation d'une partie des travaux de thèse de *Mariam Daoud* portant sur la définition et exploitation de centres d'intérêt sémantiques (Cf chapitre 2, paragraphe 2.3.2) dans un processus de recherche d'information. C'est dans le cadre de cette même thèse, que nous avons étendu nos propositions vers un cadre d'évaluation issu de *TREC-HARD*. Ce cadre d'évaluation, en cours de raffinement, nous a permis de renforcer notre participation au projet *Quaero*. En effet, les enseignements tirés, en termes de développement de stratégies de test et de simulation de contextes, nous offre des prédispositions à développer des cadres spécifiques répondant à des applicatifs imposés par le partenaire industriel *Exalead*.

5.8 Conclusion et perspectives

Nous avons cerné la question problématique de l'évaluation empirique d'un modèle d'accès contextuel à l'information et proposé une contribution permettant d'y répondre. Notre contribution dans ce sens s'inscrit dans l'approche de méthodologies d'évaluation basées sur des contextes simulés plutôt que réels et à la spécificité de réutiliser un cadre d'évaluation largement reconnu dans le domaine en l'occurrence *TREC*. Un premier cadre d'évaluation axé sur l'intégration des centres d'intérêt comme composante du modèle d'accès est le résultat de l'enrichissement d'un protocole *TREC adhoc* appliqué à une collection de test comprenant des requêtes annotées de domaines génériques. Nous avons montré que l'application d'une stratégie de test basée sur la validation croisée nous

permettait de simuler la génération de jugements de pertinence dynamique, fondée sur celles initialement formulées par les assesseurs. Un tel protocole nous a permis de nous situer par rapport à des modèles comparables dans le domaine. Un second protocole basé sur la définition de sessions de recherche nous a permis de prendre conscience de la qualité des contextes construits par la mesure de la classification de sessions de recherche selon les sujets abordés. Ce protocole nous a également permis d'identifier les meilleures mesures permettant de déterminer le degré de corrélation des sujets associés à ces sessions de recherche puis de nous situer relativement aux autres travaux connexes dans le domaine.

Les perspectives de recherche ouvertes par ce travail portent sur deux volets essentiels : définition d'un cadre d'évaluation standard pour la validation d'un accès contextuel à l'information guidé par le contexte cognitif et la prise en compte d'un contexte multi-dimensionnel dans les protocoles d'évaluation.

– *Définition d'un cadre d'évaluation standard* : une première extension de nos propositions portant sur la définition de cadres de validation d'un accès contextuel à l'information, plus particulièrement guidés par les centres d'intérêt des utilisateurs, est sans doute de passer à des évaluations menées avec des contextes réels (utilisateurs réels en situation de recherche d'information). C'est un cadre qui permettrait à court terme de renforcer nos résultats obtenus avec des contextes simulés, le cas échéant pointer sur des problèmes particuliers. Bien que ce premier travail envisageable permettrait d'élargir la portée de nos résultats, il est cependant peu généralisable. En effet, il est clair que la reproductibilité des résultats d'expérimentations de cadres de validation menés avec de vrais utilisateurs reste, à ce jour, un verrou à la fois technologique et scientifique ouvert comme en témoigne les workshops tels que *Context-based Information retrieval (CIR)*, Roskilde,2007, *Workshop on novel methodologies for evaluation in information retrieval*, Glasgow 2008, *International Workshop on Evaluating Information Access (EVIA)*, Tokyo 2008. Pour cela, notre perspective à long terme dans ce sens est d'allier les avantages d'une validation contrôlable, orientée- système et ceux d'une validation orientée-utilisateur dans le sens d'une évaluation dite *formative* (Petrelli 2008, Diaz et al. 2008), afin de définir un cadre standard permettant d'asseoir des protocoles d'évaluation de "tâches d'accès contextuel" comme le sont les tâches bien spécifiées dans les campagnes d'évaluation telles que *TREC*. De nombreuses questions sont alors ouvertes : spécification de la tâche par définition des objectifs de validation selon une approche orientée-système et/ou approche orientée-utilisateur, définition des sources permettant de construire les centres d'intérêt, définition des jalons permettant de délimiter des sessions de recherche, élaboration de mesures d'évaluation combinant l'efficacité en termes de pertinence des résultats et utilité en termes d'adéquation aux centres d'intérêt définis...

– *Prise en compte d'un contexte multidimensionnel dans le protocole d'évaluation* : les cadres définis à ce jour, sont dédiés à des technologies prenant en compte généralement une dimension unique du contexte (centre d'intérêt, lieu, ...). La prise en compte effective d'un contexte multidimensionnel dans le procédé d'évaluation est un des aspects qui permettrait à terme de développer des plate-formes d'évaluation dédiées à des applications intégrant des contextes multidimensionnels comme le commerce électronique, la formation à distance, le guide touristique etc. La grande difficulté est alors de spécifier un cadre général d'évaluation instanciable selon les contextes définis dans les applications. C'est un des moyens qui permettrait d'envisager des évaluations comparatives laissant la voie à des transferts technologiques entre domaines scientifiques faisant intervenir le contexte comme élément technologique potentiel.

Conclusion et perspectives

6

Les travaux synthétisés dans les chapitres précédents donnent un large aperçu de nos contributions à la résolution des verrous posés par l'adaptation du processus de recherche d'information à l'utilisateur et à son environnement de recherche que nous qualifions précisément de contexte. L'objectif de ce dernier chapitre est de faire l'état des lieux de nos connaissances dans le domaine, de mettre en exergue les enseignements à tirer puis de présenter les perspectives de recherche. Ces dernières constituent tantôt une synthèse (par recoupements) en complément des perspectives que nous avons annoncées dans les chapitres précédents, tantôt des ouvertures vers de nouvelles directions.

Nos investigations de recherche ont suivi le cours d'évolution des approches de RI depuis une vision orientée système vers une vision orientée contexte. Cette mutation nous a clairement guidé dans l'évolution de nos objectifs scientifiques en se posant des questions fondamentales liées à la modélisation du contexte de recherche et à la problématique sous-jacente de la clarification du besoin en information, à l'évaluation des requêtes considérant les sources d'évidence extraites du contexte ; enfin, de manière transversale, à l'évaluation empirique des modèles d'accès contextuel à l'information.

Le premier volet de nos travaux a été centré précisément autour de l'utilisateur, de son contexte, et de son besoin en information. La modélisation du contexte, étape clé de nos travaux, a été sanctionnée par :

- une stratégie d'apprentissage/évolution des centres d'intérêt de l'utilisateur : la stratégie est essentiellement basée sur la découverte de la récurrence des sujets de requête comme source d'évidence pour modéliser les centres d'intérêt. A cet effet, de nouvelles notions et mesures ont été définies. Les résultats d'expérimentations montrent l'intérêt de notre approche comparativement aux autres approches de modélisation du domaine,

- des facteurs de contenu et de qualité de source d'information permettant de décrire son profil : cette définition n'est pas absolue, elle est relativisée à l'utilisateur qui exploite la source d'information lors d'une activité de recherche d'information.

La clarification du besoin en information, rattachée à cette démarche de modélisation de l'environnement de l'utilisateur est doublement justifiée :

- la qualité de la recherche dépend de la qualité de la requête traduisant généralement son degré de clarté,
- la requête est une expression du besoin en information de l'utilisateur en mé-connaissance du contenu de l'information disponible.

L'ensemble de ces travaux consituent les briques élémentaires qui nous ont permis de mettre en place un cadre de modélisation du contexte et de la clarification des besoins en information de l'utilisateur. Cependant, de nombreuses hypothèses simplificatrices ont été posées réduisant la portée de nos propositions dans des situations de recherche réelles. Parmi les hypothèses les plus contraignantes, on citera :

1. la pertinence demeure une notion dépendante de l'adéquation du contenu du document au sujet de la requête, même s'il est relativisé au centres d'intérêt de l'utilisateur ; des critères potentiels de la pertinence situationnelle tels que la nouveauté, l'autorité, la diversité des sujets abordés par le document sont cependant en marge de nos travaux,
2. les ressources sémantiques générales à la base de définition des profils sont réutilisables pour tout domaine d'application,
3. l'historique de recherches de l'utilisateur présente, à long terme, des régularités observables permettant de garantir sa stabilité et donc la qualité des profils construits,
4. le contexte est réduit à une dimension unique,
5. une définition simple de l'intention de l'utilisateur catégorisée selon le type informationnel, type transactionnel et type navigationnel.

Il est ainsi clair que ces hypothèses nous placent en deçà des exigences des systèmes d'accès actuels. Les travaux que nous avons menés dans ce cadre nous confèrent ainsi des convictions fortes quant à l'ampleur des problèmes qui demeurent ouverts. Pour cela, une direction de recherche importante de nos travaux futurs est d'investir le thème de la clarification des besoins en information dans un contexte multidimensionnel, qui fédère actuellement de très nombreux travaux dans des disciplines très diverses. Nous envisageons de l'aborder dans sa complexité réelle à travers les exigences suivantes :

- une définition plus exhaustive de la pertinence situationnelle nous permettant de nous affranchir des barrières du sujet de la requête. On se penchera sur la définition de mesures d'utilité de l'information qui couvre divers aspects : temps, lieu, tâche, etc.
- l'intégration de nombreuses dimensions potentielles du contexte dans un cadre unifié d'accès à l'information.

En pratique, deux ouvertures marqueront nos futurs travaux. Une première ouverture, que nous investirons à court terme, portera sur des technologies de RI mobile couvrant des applicatifs caractérisés par un contexte comportant trois dimensions potentielles : les centres d'intérêt, la localisation géographique et le temps. C'est un domaine en plein expansion aujourd'hui avec l'accessibilité de sources d'information volumineuses à

travers des mobiles tels que le *e-pod* et le *GPS*[1] etc. C'est un cadre qui nous amènera à repenser globalement deux points fondamentaux :

- la représentation du contexte dans un espace spatio-temporel et la spécification de modèles d'appariement basés non plus sur la proximité thématique mais sur la proximité géographique déclinée pour un centre d'intérêt particulier. C'est une piste de recherche que nous avons commencé à explorer à travers le projet *Quaero*,

- l'exploration de nouvelles technologies pour la capture du comportement de l'utilisateur et l'inférence implicite de sa perception de la pertinence quant aux informations qu'il reçoit en situation de mobilité ; parmi les indicateurs de pertinence on citera particulièrement le mouvement des yeux et son historique de déplacements.

La seconde ouverture, que nous investirons à long terme, concerne l'introduction de la tâche à sa juste "dimension" dans le contexte et ce, dans le but de mieux adapter nos technologies à des applications potentielles telles que le domaine médical, le domaine légal, les nano-sciences etc. En effet, la littérature émanant de cogniticiens d'une part et de spécialistes en RI d'autre part, fait état d'une catégorisation plus complexe des tâches d'un utilisateur en situation de recherche d'information : résolution d'un problème, formation, contact avec autrui, exécution d'une tâche en son lieu de travail, loisir etc. Diverses taxonomies des tâches sont présentées et discutées dans (Li et Belkin 2008). Nous mènerons particulièrement des réflexions autour de l'association entre la tâche et le comportement de l'utilisateur en situation de recherche d'information. Notre objectif sera d'identifier les comportements qui caractérisent les tâches et offrir les éléments d'information qui permettent de les accomplir.

Concernant le volet de l'évaluation contextuelle des requêtes, nous nous sommes investis dans deux directions. La première porte sur l'évaluation de requêtes comportant des préférences qualitatives. Dans la continuité de notre démarche de formalisation de telles requêtes, nous avons proposé des modèles d'appariement requête-préférences-documents qui reposent essentiellement sur la proximité de graphes.
Cette contribution, loin d'être achevée, mérite d'être évaluée expérimentalement dans des contextes de recherche d'information caractérisés particulièrement par l'expression de requêtes vues comme des situations hypothétiques de recherche (caractérisées par l'imbrication de faits conditionnels) telles que le diagnostic médical, c'est une perspective que nous inscrivons à court terme dans le cadre du projet *IAPA*.
La deuxième contribution concerne la définition d'un modèle inférentiel d'accès à l'information qui met en œuvre un appariement requête-document-centres d'intérêt de l'utilisateur. Le modèle formel décrit la structure de l'information d'une part ainsi qu'un ensemble de propriétés traduisant des heuristiques en RI personnalisée, d'autre part. Les résultats de validation expérimentale menée sur des collections *TREC* montrent la viabilité de ce modèle comparativement à d'autres modèles d'accès contextuel.

[1] Global Positionning System

Ces premiers travaux autour de l'évaluation contextuelle de requête sont cependant basés sur des modèles classiques d'appariement largement adoptés dans la RI orientée système. La nouveauté a porté essentiellement sur l'introduction de la composante utilisateur dans le modèle d'appariement. Même si ces travaux nous ont effectivement permis de réviser la distribution des scores de pertinence pour un utilisateur donné, ils sont toutefois combinatoires, basés sur les hypothèses simplificatrices suivantes :

- les pertinences dérivées de la requête et du contexte de recherche sont cumulatives,

- l'utilisateur est isolé, dans le sens ou sa perception de la pertinence de l'information ne dépend pas de celle de son groupe.

Dans ce sens une piste intéressante, qu'il serait important d'investir est de poser de nouveaux fondements pour les modèles d'appariement (*ranking models*) spécifiquement adaptés à la RI contextuelle. L'usage de la stratégie d'optimisation multi-critères telles qu'elle est abordée par les AG's, serait à notre avis une voie intéressante pour passer outre la première hypothèse.
Sous un autre angle, la prise en compte, dans le modèle d'appariement, de *l'autorité* des individus en qualité d'auteurs ou d'utilisateurs ainsi que leur position dans le réseau social, permettrait de lever la seconde hypothèse. Dans ce dernier cas, une large ouverture est donnée pour l'utilisation de la topologie des réseaux sociaux pour la mise en place de modèles d'ordonnancement de documents ; ces modèles seront basés sur la prise en compte intrinsèque du contenu des documents d'une part et d'autre part, de la position sociale, inférée dans le réseau, de leur auteur (personne ayant produit le document) ou utilisateurs (personnes ayant exprimé le besoin, jugé la pertinence).

Concernant la problématique de l'évaluation de l'efficacité des modèles d'accès contextuel à l'information, nos choix de solutions ont été guidés par nos contributions passées dans le domaine. En effet, notre premier objectif dans cadre étant de valider l'efficacité d'un modèle d'accès intégrant les centres d'intérêt de l'utilisateur, nous avons ciblé un protocole d'évaluation qui vise l'augmentation du cadre d'évaluation *TREC* par une telle ressource. Nous avons opté pour le principe de simulation des utilisateurs pour minimiser le cout de réalisation de nos expérimentations. Pour éviter le biais de l'évaluation, nous avons également défini une stratégie de test par validation croisée.
Ce cadre, préliminaire, nous a permis de valider nos propositions dans le domaine, d'identifier les problèmes et de proposer des solutions. Cependant, le paradigme sous-jacent de l'évaluation est intrinsèquement "fermé" en ce sens que :

1. les requêtes ont des descriptions riches, peu comparables aux requêtes des utilisateurs du web par exemple,

2. les centres d'intérêt, même simulés par le principe de validation croisée, émanent à la base d'assesseurs qui expriment une pertinence thématique,

3. de nombreux aspects liés à l'utilisateur telles que sa familiarité avec le sujet de la requête, son but, son expertise, ne sont pas modélisables dans un modèle d'évaluation de type laboratoire comme celui que nous avons spécifié.

Pour de nombreuses autres raisons encore liées à la définition de variables d'expérimentation dépendantes du contexte, il est clair, que l'évaluation de l'accès contextuel à l'information demeurera à notre avis un problème ouvert pour la prochaine décennie. A ce titre, nous envisageons de rassembler des membres de la communauté pour un échange d'expériences autour de ce problème par le biais d'un *workshop* annexé à la conférence de référence du domaine en l'occurrence *ECIR* qui aura lieu à *Toulouse* en avril 2009. Nous inscrivons le thème de l'évaluation dans une direction prioritaire de nos recherches à court terme que nous étalerons sur une longue période. Dans ce sens, nous participons actuellement activement à l'évaluation de la tâche 2.6. dans le cadre du projet *Quaero*. Un partenariat monté avec la société *Exalead* nous permettrait à court terme de spécifier un protocole d'évaluation en mode *batch* basé sur les contenus de fichiers *logs* d'utilisateurs d'un moteur de recherche sur le web. Nous poursuivrons nos investigations à travers des campagnes d'évaluation que nous allons mener avec de vrais utilisateurs, également dans le projet *Quaero*.
Notre objectif à long terme est de spécifier des cadres d'évaluation, à l'instar de ceux proposés à *TREC*, offrant des ressources partageables destinées à être exploitées pour l'évaluation d'un accès contextuel à l'information sur un environnement ouvert tel que le web, et ce, indépendamment de toute technologie. C'est l'une de nos ambitions majeures à travers le projet *Quaero*.

Enfin, en complément de l'évaluation de l'efficacité, il nous est important de signaler, à l'issue de l'ensemble de nos travaux, que nous n'avons pas explicitement abordé l'efficience des modèles produits. Nous pouvons affirmer que les algorithmes mis en œuvre passent l'échelle des collections *TREC* (volumineuses) que nous avons utilisées lors de notre validation expérimentale. En revanche, nous ne sommes pas en mesure aujourd'hui d'avoir un ordre de grandeur de leurs performances, en termes d'efficience, sur des collections plus volumineuses telles que celles disponibles, le cas extrême, sur le web. Nous pensons particulièrement à trois principaux facteurs de calcul induits par nos technologies :

- l'exécution des procédures d'appariement dans le cas de la définition de profils sémantiques qui utilisent une ontologie de référence,
- l'exécution de procédures d'évaluation de requête basées sur les CP-Nets,
- l'exécution parallèle d'AG's pour l'optimisation de requêtes.

Pour cela, nous envisageons tout d'abord de porter nos solutions sur un cadre adéquat qui offre les ressources nécessaires pour permettre cette évaluation, en l'occurrence la plate-forme OSIRIM (http ://www.irit.fr/OSIRIM). C'est un support technologique qui consiste à fédérer des ressources distribuées (corpus d'évaluation, outils d'analyse) sur une même plateforme pour expérimenter des chaînes complexes de traitement permettant de retrouver des contenus textuels, audio et vidéo pertinents. C'est un cadre qui nous permettra, en premier lieu, de rendre nos solutions accessibles pour la communauté et notre cadre d'évaluation réutilisable pour d'autres technologies ; ceci constitue en soi, une retombée scientifique importante de nos travaux.
Au delà, nous serons éventuellement confrontés à la question épineuse du passage à l'échelle largement posée dans le domaine. Nous envisageons d'y contribuer en mettant à l'épreuve les solutions déjà proposées dans le cas de modèles d'accès classique, faisant partie d'une expertise approuvée au sein de notre composante, vers des modèles augmentés par le contexte.

Annexes

A

Bibliographie

E. Agichtein, E. Brill, et S. T. Dumais. Improving web search ranking by incorporating user behavior information. Dans *Proceedings of the* 29^{th} *International SIGIR conference on research and development in information retrieval*, pages 19–26, 2006. (Cité pages 9, 46, 70 et 89.)

J. Allan. Hard track overview in trec 2003 high accuracy retrieval from documents. Dans *Proceedings of the* 12^{th} *text retrieval conference (TREC-12)*, pages 24–37. National Institute of Standards and Technology, NIST special publication, 2003a. (Cité pages 24 et 88.)

J. Allan. Hard track overview in trec 2003 : High accuracy retrieval from documents. Dans *TREC*, pages 24–37, 2003b. (Cité page 97.)

B.L. Allen. Cognitive research in information science : Implications for design. *Annual Review of Information Science and Technology (ARIST)*, 26 : 3–37, 1991. (Cité page 4.)

B. Anand, S.S.and Mobasher. Introduction to intelligent techniques for web personalization. *ACM Transactions on Internet Technologies*, 7(4) :18, 2007. ISSN 1533-5399. (Cité pages 15, 29 et 42.)

C. Anderson, P. Domingos, et D. Weld. Personalizing web sites for mobile users. Dans *C. R. Anderson, P. Domingos, and D. S. Weld. Personalizing web sites for mobile users. In Proceedings of the* 10^{th} *International WWW Conference*, pages 565–575, 2001. (Cité page 46.)

F. Bacchus et A.J. Grove. Graphical models for preference and utility. Dans *Proceedings of the* 11^{th} *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 3–10, august 1995. (Cité page 58.)

R. Baeza-Yates, L. Calderon-Benavides, et Gonzalez-Caro. The intention behind web queries. Dans *Proceedings of String Processing and Information REtrieval(SPIRE 2006)*, pages 98–109, 2006. (Cité page 52.)

R. Baeza-Yates et B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999. ISBN 0-201-39829-X. URL `citeseer.ist.psu.edu/baeza-yates99modern.html`. (Cité pages 3 et 29.)

J. Bai, J. Nie, H. Bouchard, et G. Cao. Using query contexts in information retrieval. Dans *Proceedings of the 31 st annual internationsl ACM SIGIR Conference on Research and development in Information retrieval*, pages 15–22, 2007. (Cité pages 71 et 103.)

M. Balabanović et Y. Shoham. Content-based collaborative recommendation. *Commununications of the ACM*, 40(3) :66–72, 1997. ISSN 0001-0782. (Cité page 68.)

M. Bazire et P. Brézillon. *Understanding Context Before Using It*. 2005. URL `http://dx.doi.org/10.1007/11508373_3`. (Cité page 7.)

M. Baziz, M. Boughanem, et N. Aussenac-Gilles. A conceptual indexing approach based on document content representation. Dans F. Crestani et I. Ruthven, éditeurs, *CoLIS5 : Fifth International Conference on Conceptions of Libraries and Information Science , Glasgow, UK, 04/06/05-08/06/05*, pages 171–186, Berlin Heidelberg, juin 2005a. Lecture Notes in Computer Science LNCS Volume 3507/2005, Springer-Verlag. URL `http://dx.doi.org/10.1007/11495222_14`. (Cité page 49.)

M. Baziz, M. Boughanem, et N. Aussenac-Gilles. Evaluating a Conceptual Indexing Method by Utilizing WordNet . Dans Carol Peters, Fredric C. Gey, Julio Gonzalo, et Gareth J.F. Jones, éditeurs, *Accessing Multilingual Information Repositories : 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Revised Selected Papers , Vienna, Austria, 21/09/05-23/09/05*. Lecture Notes in Computer Science, Vol. 4022, septembre 2005b. (Cité page 91.)

M. Beaulieu. Experiments with interfaces to support query expansion. *Journal of documentation*, 53(1) :8–19, 1997. (Cité page 49.)

I. M. Begg, J. Gnocato, et W. E. Moore. A prototype intelligent user interface for real-time supervisory control systems. Dans *Proceedings of the 1st international conference on Intelligent user interfaces*, pages 211–214, New York, NY, USA, 1993. ACM. (Cité page 31.)

N.J Belkin. Information concepts for information science. *Journal of documentation*, 34 :55–85, 1978. (Cité pages 2 et 4.)

NJ. Belkin. Knowledge elicitation using discourse analysis. *International Journal of Man-Machine studies*, 27 :127–144, 1987. (Cité page 4.)

N.J. Belkin. Interaction with texts : Information retrieval as information seeking behavior. Dans *Proceedings of Information Retrieval Conference*, pages 55–66, 1993. (Cité page 50.)

N.J Belkin, C. Cool, J. Koenemann, S. Park, et W.B Ng. Information seeking behaviour in new searching environments. Dans *Ingwersen, P., Pors*, pages 403–416, 1996. (Cité page 5.)

N.J. Belkin et W.B. Croft. Information filtering and information retrieval : two sides of the same coin ? *Communications of the ACM*, 35(12) :29–38, 1992. ISSN 0001-0782. (Cité page 5.)

H. Beorko et M. Bernick. Automatic document classification. *Journal of the Association for Computing Machinery*, pages 151–161, 1963. (Cité page 3.)

D. Bilal. Children's use of the yahooligans ! web search engine : cognitive, physical, and affective behaviors on fact-based search tasks. *Journal of American Society in Information Science*, 51(7) :646–665, 2000. (Cité page 8.)

A. Bookstein. An approach to weighted boolean searches. *Journal of the American Society for Information Science (JASIS)*, 31(4) :240–247, 1980. (Cité page 51.)

G. Bordogna, P. Carrara, et G. Pasi. Query term weights as constraints in fuzzy information retrieval. *Information Processing and Management*, 27 (1) :15–26, 1991. ISSN 0306-4573. (Cité pages 20, 21, 50, 51, 63 et 69.)

G. Bordogna et G. Pasi. Linguistic aggregation operators of selection criterian fuzzy information retrieval. *International Journal of Intelligent Systems*, 10(2) :233–248, 1991. (Cité pages 50, 51, 63 et 70.)

G. Bordogna et G. Pasi. A fuzzy linguistic approach generalizing boolean information retrieval. *Journal of the American Society for Information Science*, 44(2) :15–26, 1999. (Cité page 47.)

G. Bordogna et G. Pasi. Linguistic aggregation operators of selection criteria in fuzzy information retrieval. *International Journal of Intelligent Systems*, 10(2) :233–248, 2007. (Cité page 21.)

P. Borlund. The concept of relevance in ir. *Journal of the American Society for Information Science and Technology*, 54(10) :913–925, 2003a. (Cité pages 5 et 87.)

P. Borlund. The iir evaluation model : A framework for evaluation of interactive information retrieval systems. *Journal of Information Research*, 8(3) :152, 2003b. (Cité pages 5, 23 et 89.)

F. Boubekeur. *Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets*. Thèse de Doctorat en Informatique, Université Paul Sabatier (Toulouse), 2008. (Cité pages 59 et 74.)

F. Boubekeur, M. Boughanem, et L. Tamine. Towards flexible information retrieval based on cp-nets. Dans Henrik Legind Larsen, Gabriella Pasi, et Daniel Ortiz-Arroyo, éditeurs, *Flexible Query Answering (FQAS), Milan, Italie, 07/01/2006-10/06/2006*, Advances in Artificial Intelligence, pages 222–231, http ://www.wspc.com.sg/, juin 2006. World Scientific Publishing. (Cité pages 26, 71, 72 et 82.)

F. Boubekeur, M. Boughanem, et L. Tamine. Semantic information retrieval based on cp-nets,. Dans *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), London, 23/07/07-26/07/07*, page (electronic medium), http ://www.ieee.org/, juillet 2007. IEEE. (Cité pages 26, 73, 75 et 82.)

F. Boubekeur, M. Boughanem, et L. Tamine. Exploiting association rules and ontology for semantic document indexing. Dans *International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU), Malaga, 22/06/08-27/06/08*, juin 2008. (Cité pages 26, 73 et 82.)

M. Boughanem, C. Chrisment, et Ch. Soulé-Dupuy. Query modification based on relevance back-propagation in adhoc environnement. *Information Processing and Management*, 35 :121–139, avril 1999. (Cité page 49.)

M. Boughanem, C. Chrisment, et L. Tamine. On using genetic algorithms for multimodal relevance optimisation in information retrieval. *Journal of American Society in Information Systems and Technology (JASIST)*, 53(11) : 934–942, 2002a. (Cité pages 26 et 63.)

M. Boughanem, Y. Loiseau, et H. Prade. Rank-ordering documents according to their relevance in information retrieval using refinements of ordered-weighted aggregations. Dans *In Proceedings of the 3rd International Workshop on Adaptive Multimedia Retrieval (AMR'05)*, pages 44–54, 2005. (Cité pages 21 et 70.)

M. Boughanem, Y. Loiseau, et H. Prade. Refining aggregation functions for improving document ranking in information retrieval. Dans *International Conference on Scalable Uncertainty Management (SUM 2007), Washington,DC, USA, 10/10/07-12/10/07*, volume 4772/2007, pages 255–267, http ://www.springerlink.com/, octobre 2007. Springer-Verlag. (Cité page 22.)

M. Boughanem, H. Tebri, et M. Tmar. Irit at trec 2002 : Filtering track. Dans *The Eleventh Text Retrieval Conference (TREC 2002)*, Gaithersburg, Maryland(USA), novembre 2002b. Text Retrieval Conference, TREC. (Cité pages 68 et 91.)

C. Boutilier, F. Bacchus, et R. Brafman. Ucp-networks : A directed graphical representation of conditional utilities. Dans *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 56–64, 2001. (Cité page 58.)

C. Boutilier, R. Brafman, H. Hoos, et D. Poole. Reasoning with conditional ceteris paribus preference statements. Dans *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 71–80, 1999. (Cité page 57.)

M. Bouzeghoub et V. Peralta. A framework for analysis of data freshness. Dans *Proceedings of the 1st Workshop on Information Quality in Information Systems (IQIS)*, 2004. (Cité pages 32 et 46.)

B.C Brookes. The developping cognitive viewpoint in information science. Dans *De Mey, M. and al Editors*, pages 195–203, septembre 1977. (Cité page 4.)

C. Buckley, G. Salton, J. Allan, et A. Singhal. Automatic query expansion using smart :trec-3. Dans *Proceedings of the 3rd Text REtrieval Conference*, pages 69–80, 1994. (Cité page 49.)

C. Buckley, A. Singhal, M. Mandar, et G. Salton. New retrieval approaches using smart : Trec 4. Dans *Proceedings of the 4th Text REtrieval Conference*, pages 25–48, 1995. (Cité page 49.)

J. Budzik et K.J. Hammond. User interactions with every day applications as context for just-in-time information access. Dans *Proceedings of the 5th international conference on intelligent user interfaces*, pages 44–51, Mars 2000. (Cité page 18.)

D. Buell et D. Kraft. A model for a weighted retrieval system. *Journal of the American Society for Information Science (JASIS)*, 32(3) :211–216, 1981a. (Cité page 51.)

D. Buell et D. Kraft. Threshold values and boolean retrieval systems. *Information Processing and Management (IPM)*, 17(3) :127–136, 1981b. (Cité page 69.)

K. Bystrom et K. Jarvelin. Task complexity affects information seeking and use. *Information Processing and Management*, 31(2) :191–213, 1995. (Cité page 89.)

J. Callan, Z. Lu, et W.B. Croft. Searching distributed collections with inference networks. Dans *Proceedings of the* 18^{th} *ACM SIGIR International Conference on Research and Development*, pages 21–28, 1995. (Cité page 32.)

S. Cater et D. Kraft. A generalization and classification of the waller-kraft wish-list. *Information Processing and Management (IPM)*, 25(15) :15–25, 1989. (Cité page 69.)

H. Chen. Machine learning for information retrieval : Neural networks, symbolic learning and genetic algorithms. *Journal of the American Society for Information Science and Technology*, 46(3) :194–216, 1995. (Cité pages 53 et 62.)

L. Chen et K. Sycara. Webmate : A personal agent for browsing and searching. Dans *Proceedings of the* 2^{nd} *international conference on autonomous agents and multi agent systems, Minneapolis*, pages 10–13, 1998. (Cité page 31.)

K. Cheverst, N. Davies, K. Mitchell, A. Friday, et C. Efstratiou. Developing a context-aware electronic tourist guide : some issues and experiences. Dans *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 17–24, New York, NY, USA, 2000. ACM. (Cité pages 9, 47, 51 et 70.)

D.N. Chin. Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction*, 11(1-2) :181–194, 2001. ISSN 0924-1868. (Cité page 89.)

P. Chirita, D. Olmedilla, et W. Nejdl. Pros : a personalized ranking platform for web search. Dans The Netherlands Springer, éditeur, *Proceedings of the* 3^{rd} *International Conference Adaptaive Hypermedia and Adaptive Web-based Systems*, pages 34–43. Lecture Notes in Computer Science, Vol. 3137, august 2004. URL `citeseer.ist.psu.edu/chirita04pros.html`. (Cité page 68.)

L. Chittaro, éditeur. *Human-Computer Interaction with Mobile Devices and Services*, volume 2795 de *Lecture Notes in Computer Science*, 2003. Springer. ISBN 3-540-40821-5. (Cité page 46.)

K. Church, W. Gale, P. Hanks, et D. Hindle. Parsing, word associations and typical predicated-argument relations. Dans *Proceedings of the 1989 DARPA speech and natural language Workshop*, 1989. (Cité page 4.)

C.W Cleverdon. The cranfield test on index language devices. Dans *Aslib*, pages 173–194, 1967a. (Cité pages 4, 12, 85 et 86.)

C.W Cleverdon. The cranfield test on index language devices. Dans *Aslib*, pages 173–194, 1967b. (Cité page 23.)

C. Cool et A. Spink. Issues of context in information retrieval : an introduction to the special issue. *In Journal of Information Processsing and Management (IPM)*, 38(55) :605–611, 2002. (Cité page 7.)

F. Crestani. Exploiting the similarity of non-matching terms at retrieval time. *Information Retrieval*, 2(1) :27–47, 2000. (Cité page 47.)

W. Croft et D. Harper. Using probabilistic models for information retrieval without relevance information. *Journal of documentation*, 35(4) :285–295, 1979. (Cité page 49.)

M. Daoud, L. Tamine, et M. Boughanem. Using a concept-based user context for search personalization. Dans *International Conference of Data Mining and Knowledge Engineering (ICDMKE), London, 02/07/2008-04/07/2008*, 2008a. (Cité pages 25, 27, 36, 37, 71 et 104.)

M. Daoud, L. Tamine, et M. Boughanem. Using a graph-based ontological user profile for personalizing search. Dans *Proceedings of the* 17^{th} *International Conference and Knowledge Management, CIKM 2008, washington, 06/12/08-11/11/08*. Arvin Agah, Jamie Callan, Elke Rundensteiner, novembre 2008b. (Cité page 27.)

M. Daoud, L. Tamine-Lechani, et M. Boughanem. Learning user interests for a session-based personalized search. Dans *Proceedings of the the Second International IIiX Symposium (IIIX'08). To appear*, pages 293–298. IAENG, 2008c. (Cité pages 9, 25, 27, 36, 44, 71 et 104.)

M. Daoud, L. Tamine-Lechani, M. Boughanem, et B. Chebaro. Learning implicit user interests using ontology and search history for personalization. Dans *InternationalWorkshop on Personalized Access toWeb Information (PAWI) within the* 8^{th} *International Web Information Systems Engineering (WISE), Nancy, 03/12/2007-07/12/2007*, décembre 2007. (Cité pages 25 et 36.)

N. Davies, K. Mitchell, K. Cheverest, et G. Blair. Developing a context sensitive tourist guide. Dans *First Workshop on Human Computer Interaction with Mobile Devices*. GIST Technical Report G98-1, 1998. (Cité page 6.)

M. De Mey. The cognitive viewpoint : Its development and its scope. *International Workshop on the cognitive viewpoint*, pages 285–295, 1977. (Cité page 2.)

S. Deerwester, S.T. Dumais, G.W. Furnas, T. Landauer, et Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6) :391–407, 1990. (Cité page 47.)

B. Dervin et M. Nilan. Information needs and uses. *ARIST, William, M.E. Eds*, pages 3–33, 1986. (Cité page 4.)

A. Diaz, A. Garcia, et P. Gervas. User-centred versus system-centred evaluation of a personalization system. *Information Processing and Management*, 44(3) :1293–1307, 2008. ISSN 0306-4573. (Cité pages 89 et 105.)

C. Ding et J. C. Patra. User modeling for personalized web search with self-organizing map. *Journal of American Society in Information Science and Technology*, 58(4) :494–507, 2007. ISSN 1532-2882. (Cité pages 24, 46 et 89.)

D. Dubois et H. Prade. Review of fuzzy sets aggregation connectives. *Information Sciences*, 3 :85–121, 1985. (Cité page 70.)

D. Dubois et H. Prade. Weighted minimum and maximum operations in fuzzy set theory. *Information Sciences*, 39 :205–210, 1986. (Cité pages 21 et 69.)

S. Dumais, E. Cuttrel, J.J. Cadiz, G. Jancke, R. Sarin, et D.C Robbins. Stuff i've seen : a system for a personal information retrieval and re-use. Dans *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development*, pages 72–79, Toronto, Canada, 2003a. ACM Press. (Cité page 70.)

S. Dumais, Cuttrell E., J.J. Cadiz, G. Jancke, R. Sarin, et D.C. Robbins. Stuff i've seen : a system for a personal information retrieval and re-use. Dans *Proceedings of the* 26^{th} *ACM SIGIR*, pages 72–79. Toronto, Juillet 2003b. (Cité pages 18 et 31.)

B. Duncan et D. H. Kraft. (Cité page 20.)

E.N Efthimiadis. Query expansion. *Annual Review of Information Science and Technology (ARIST)*, 31 :121–187, 1996. (Cité page 47.)

W. Fan, M.D Gordon, et P. Pathak. Discovery of context specific ranking functions for effective information retrieval using genetic programming. volume 16, pages 523–527, 2004. (Cité page 70.)

H. Fargier et P. Perny. A characterisation of generalised concordance rules in multicriteria decision making. *In International Journal of Intelligent Systems (IJIS)*, 18 :751–774, 2003. (Cité page 81.)

E. Frias-Martinez, S.Y. Chen, R. D. Macredie, et X. Liu. The role of human factors in stereotyping behavior and perception of digital library users : a robust clustering approach. *User Modeling and User-Adapted Interaction*, 17(3) :1573–1391, 2007. ISSN 0924-1868. (Cité page 8.)

N. Fuhr. A decision-theoretic approach to database selection in networked ir. *ACM Transactions on Information Systems*, 17(3) :229–249, 1999. (Cité page 32.)

S. Gauch, J. Chaffee, et Pretschnerm A. Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems*, 1(3-4) :219–234, 2003. (Cité pages 17, 18, 22, 31, 70, 71 et 82.)

G. Gentili, A. Micarelli, et F. Sciarrone. Infoweb : An adaptive information filtering system for the cultural heritage domain. *Applied Artificial Intelligence*, 17(8-9) :715–744, 2003. (Cité page 31.)

E. Glover, G.W. Flake, S. Lawrence, W.P. Birmingham, A. Kruger, C.L. Giles, et D.M. Pennock. Improving category specific web search by learning query modifications. Dans *Proceedings of Symposium on Applications and the Internet*, pages 23–31, January 2001. (Cité page 32.)

A. Göker et H. Myrhaug. Evaluation of a mobile information system in context. *Information Processing and Management*, 44(1) :39–65, 2008. ISSN 0306-4573. (Cité page 7.)

A. Göker et H. I. Myrhaug. User context and personalisation. Dans *ECCBR Workshop on Case Based Reasoning and Personalisation, Aberdeen*, 2002. (Cité page 7.)

A. Göker, S. Watt, H.I. Myrhaug, N. Whitehead, M. Yakici, R. Bierig, S. Kanth, et H. Cumming. An ambient, personalised, and context-sensitive information system for mobile users. Dans *Proceedings of the* 2^{nd} *European Union symposium on Ambient intelligence (EUSAI)*, pages 19–24, New York, NY, USA, 2004. ACM. ISBN 1-58113-992-6. (Cité pages 47 et 70.)

M. D. Gordon. User-based document clustering by redescribing subject descriptions with a genetic algorithm. *Journal of the American Society for Information Science and Technology*, 42(5) :311–322, 1991. (Cité page 53.)

L. Gravano, C. Chang, H. Garcia-Molina, et A. Paepcke. Starts : Stanford proposal for internet metasearching. Dans *Proceedings of the* 20^{th} *ACM-SIGMOD International Conference on Management of Data*, pages 207–218, 1997. (Cité page 32.)

H. Haddad. French noun phrase indexing and mining for an information retrieval system. Dans *Proceedings of the* 13^{rd} *Conference on String Processing and Information REtrieval (SPIRE)*, pages 277–286, 2003. (Cité page 49.)

D. Harman. Towards interactive query expansion. Dans *Proceedings of the* 11^{th} *annual international ACM SIGIR conference on Research and development in information retrieval*, pages 321–331, 1988. (Cité page 20.)

D. Harman. *Relevance feedback and other query modification techniques*. In W.B. Frakes and R. Baeza-Yates, editors, Information Retrieval : data structures and algorithms, chapter 11, Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992a. (Cité pages 47 et 49.)

D.K. Harman. Overview of the the 1^{st} text retrieval conference (trec-1). Dans *Proceedings of the* 1^{st} *text retrieval conference (TREC-1)*, pages 1–20. National Institute of Standards and Technology, NIST special publication, 1992b. (Cité page 4.)

D.K Harman. Overview of the the 1st text retrieval conference (trec-4). Dans *Proceedings of the* 1^{st} *text retrieval conference (TREC-4)*, pages 1–24. National Institute of Standards and Technology, NIST special publication, 1995a. (Cité page 87.)

D.K Harman. Overview of the the 4^{th} text retrieval conference (trec-4). Dans *Proceedings of the 4^{th} text retrieval conference (TREC-4)*, pages 1–24. National Institute of Standards and Technology, NIST special publication, 1995b. (Cité page 24.)

S. Hattori, T. Tezuka, et K. Tanaka. Context-aware query refinement for mobile web search. Dans *Proceedings of the 2007 International Symposium on Applications and the Internet Workshops*, page 15, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-2757-4. (Cité page 46.)

T. Haveliwala. Topic-sensitive page rank. Dans *International ACM World Wide Web conference*, pages 727–736, 2002a. (Cité page 9.)

T. H. Haveliwala. Topic-sensitive pagerank. Dans *Proceedings of the Eleventh International World Wide Web Conference (WWW 2002)*, Honolulu, Hawaii, May 2002b. (Cité page 68.)

T. Hofmann. Probabilistic latent semantic analysis. Dans *Proceedings of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999. URL `http://citeseer.ist.psu.edu/hofmann99probabilistic.html`. (Cité page 70.)

J. Holland. *Adaptation In Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975. (Cité page 52.)

M E. Hupfer et B. Detlor. Gender and web information seeking : A self-concept orientation model. *Journal of the American Society for Information Science and Technology*, 57(8) :1105 – 1115, 2006. (Cité page 8.)

P. Ingwersen. Polyrepresentation of information needs and semantic entities : elements of a cognitive theory for information retrieval interaction. Dans *Proceedings of the 17^{th} annual international ACM SIGIR conference on Research and development in information retrieval*, pages 101–110, New York, NY, USA, 1994a. Springer-Verlag New York, Inc. ISBN 0-387-19889-X. (Cité pages 9, 50 et 53.)

P. Ingwersen. Polyrepresentation of information needs and semantic entities : Elements of a cognitive theory of information retrieval and interaction. Dans *Croft W.B, Van Risjsbergen, C.J Eds*, pages 101–111. Proceedings of the 17^{th} ACM SIGIR International Conference on Research and Development, Août 1994b. (Cité pages 5 et 64.)

P. Ingwersen. Cognitive perspectives of information retrieval interaction : Elements of a cognitive theory. *Journal of documentation*, 52(1) :3–50, 1996a. (Cité pages 4 et 7.)

P. Ingwersen. Cognitive perspectives of information retrieval interactions : Elements of a cognitive ir theory. *Annual review of information science and technology*, 52(1) :3–50, 1996b. (Cité page 4.)

P. Ingwersen et K. Jarvelin. *The turn : Integration of information seeking and retrieval in context*. Springer, 2005. (Cité pages 4, 29 et 85.)

Y. E. Ioannidis et G. Koutrika. Personalized systems : Models and methods from an ir and db perspective. Dans *VLDB*, page 1365, 2005. (Cité page 6.)

R. Iqbal, J. Sturm, O. Kulyk, J. Wang, et J. Terken. User-centred design and evaluation of ubiquitous services. Dans *Proceedings of the* 23^{rd} *annual international conference on Design of communication (SIGDOC '05)*, pages 138–145, New York, NY, USA, 2005. ACM. ISBN 1-59593-175-9. (Cité page 70.)

B. J. Jansen, D. L. Booth, et A. Spink. Determining the user intent of web search engine queries. Dans *Proceedings of the* 16^{th} *international conference on World Wide Web*, pages 1149–1150, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. (Cité pages 9, 21, 46 et 51.)

B.J. Jansen, A. Spink, et V. Kathuria. How to define searching sessions on web search engines. Dans *In WEBKDD'06*, pages 92–109, 2006. (Cité page 31.)

K. Jarvelin et J. Kekalainen. Cumulative gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (ACM TOIS)*, 20(4) : 422–446, 2002. (Cité page 89.)

G. Jeh et J. Widom. Scaling personalized web search. Dans *Proceedings of the* 12^{th} *international conference on World Wide Web*, pages 271–279, New York, NY, USA, 2002. ACM Press. ISBN 1-58113-680-3. (Cité pages 46 et 68.)

F.V Jensen. Bayesian networks and decision graphs. Springer, 2001. (Cité pages 77 et 79.)

T. Joachims, L. Granka, H. Hembrooke, F. Radlinski, et G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information systems*, 25(2) :7, April 2007. (Cité pages 9, 18, 31 et 46.)

K. Järvelin. On information, information technology and the development of society : An information science perspective. Dans *Ingwersen, P., Kajberg, L. and Pejtersen, A.M Editors*, pages 35–55, 1986. (Cité page 4.)

In-Ho Kang et G. Kim. Query type classification for web document retrieval. Dans *Proceedings of the* 26^{th} *annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 64–71, New York, NY, USA, 2003. ACM. ISBN 1-58113-646-3. (Cité pages 21, 46, 51, 52 et 64.)

P.B. Kantor. Information retrieval techniques. *Annual Review of Information Science and Technology (ARIST)*, 29 :53–90, 1994. (Cité page 4.)

S. Kechid, H. Drias, et L. Tamine. Personnalized access to multiple information sources. Dans Vicente Guerrero, éditeur, *International Conference on Multidisciplinary Sciences and Technologies (InSciT), Merida, Spain, 25/10/2006-28/10/2006*, pages 546–552, http ://www.instac.es, octobre 2006. Open Institute of Knowledge. (Cité page 25.)

S. Kechid, L. Tamine, et H. Drias. Personalizing information retrieval in a distributed environment. *International Review on Computers and Software*, 2(2) :98–107, mars 2007. (Cité page 25.)

J. Kekalainen et K. Jarvelin. Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. Dans *Proceedings of the* 4^{th} *CoLIS conference*, pages 253–270. P. Ingwersen and P. Vakkari, 2004. (Cité pages 23 et 86.)

D. Kelly et X. Fu. Eliciting better information need descriptions from users of information search systems. *Information Processing and Management*, 43(1) :30–46, 2007. (Cité page 18.)

N. J Kelly. Understanding implicit feedback and document preference : a naturalistic study. Dans *PHD dissertation*. Ritgers University, Etats-unis, Janvier 2004. (Cité page 51.)

K. Kim et B. Allen. Cognitive and task influences on web searching behavior. *Journal of American Society in Information Science and Technology*, 53 (2) :109–119, 2002. ISSN 1532-2882. (Cité page 87.)

K.S. Kim. Effects of emotion control and task on web searching behavior. *Information Processing and Management*, 44(1) :373–385, 2008. (Cité page 8.)

A. Kobsa. User modeling in dialog systems : Potentials and hazards. *AI and Society*, 4 :214–240, 1990. (Cité page 13.)

A. Kobsa et W. Wahlster. *User Models in Dialog Systems*. Springer-Verlag, 1989. (Cité page 10.)

D. Kostadinov, M. Bouzeghoub, et S. Lopes. Accès personnalisé à des sources de données multiples : évaluation de deux approches de reformulation de requêtes. *Revue Ingénierie des Systèmes d'Information (ISI)*, 2008. (Cité page 6.)

G. Koutrika et Y.E. Ioannidis. Constrained optimalities in query personalization. Dans *SIGMOD Conference*, pages 73–84, 2005. (Cité page 46.)

FE. Buckles BP. Kraft, DH. Petry et T. Sadisavan. *Applying genetic algorithms to information retrieval system via relevance feedback*. Physica Verlag, Heidelberg, Germany, 1995. (Cité pages 53 et 62.)

K. Lang. NewsWeeder : learning to filter netnews. Dans *Proceedings of the* 12^{th} *International Conference on Machine Learning*, pages 331–339. Morgan Kaufmann publishers Inc. : San Mateo, CA, USA, 1995. (Cité page 9.)

B. Larsen, H. Lund, J.K. Andreasen, et P. Ingwersen. Using value-added document representations in inex. Dans *INEX Workshop proceedings*, pages 67–72, 2003. (Cité page 5.)

E. Law, T. Klobucar, et M. Pipan. User effect in evaluating personalized information retrieval systems. Dans *EC-TEL*, pages 257–271. Springer-Verlag, 2006. (Cité pages 23, 86 et 87.)

J. Lee, X. Hu, et J.S. Downie. Qa websites : Rich research resources for contextualizing information retrieval behaviors. Dans *Proceedings of the* 28^{th} *International SIGIR conference on research and development in information retrieval, Workshop on information retrieval in context*, pages 33–366, 2005a. (Cité pages 24, 46, 70 et 89.)

J.H. Lee. Combining the evidence of different relevance feedback methods for information retrieval. *Information Processing and Management (IPM)*, 34(6) :681–691, 1998. (Cité page 50.)

U. Lee, Z. Liu, et J. Cho. Automatic identification of user goals in web search. Dans *Proceedings of the* 14^{th} *international conference on World Wide Web*, pages 391–400, New York, NY, USA, 2005b. ACM. ISBN 1-59593-046-9. (Cité pages 21 et 51.)

M.E Lesk. Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from a nice cream cone. Dans *Proceedings of the ACM Special Interest Group for Design of Communications (SIGDOC) Conference*, LNCS, pages 24–26. ACM, 1986. (Cité page 47.)

A. Leuski. Context features in email archives. Dans *Proceedings of the* 28^{th} *International SIGIR conference on research and development in information retrieval, Workshop on information retrieval in context*, 2005. (Cité page 46.)

Y. Li et N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Inf. Process. Manage.*, 44(6) :1822–1837, 2008. ISSN 0306-4573. (Cité page 109.)

Y. Li, R. Krishnamurthy, Vaithyanathan Sh., et H.V. Jagadish. Getting work done on the web : supporting transactional queries. Dans *Proceedings of the* 29^{th} *annual international ACM SIGIR conference on Research and development in information retrieval*, pages 557–564, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. (Cité pages 21 et 52.)

H. Lieberman. Letizia : An agent that assists web browsing. Dans *Proceedings of the* 14^th *International Joint Conference On Artificial Intelligence*, Montreal, Canada, August 1995. (Cité pages 5, 18, 46 et 68.)

C. Lin, G.R. Xue, H.J. Zeng, et Y. Yu. Using probabilistic latent semantic analysis for personalized web search. Dans *PWeb Conference, Springer Verlag Eds*, pages 707–711, 2005a. (Cité pages 22, 46 et 70.)

C. Lin, G.R Xue, H.J Zeng, et Y. YU. Using probabilistic latent semantic analysis for personalized web search. Dans *Proceedings of the APWeb Conference*, pages 707–717, Berlin Heidelberg, 2005b. Springer-Verlag. (Cité pages 70, 71 et 82.)

S.H. Lin, C.S Shih, M.C Chen, J. Ho, M. Ko, et Y.M. Huang. Extracting classification knowledge of internet documents with mining term-associations : A semantic approach. Dans *In the* 21^{th} *International SIGIR Conference on Research end Development in Information Retrieval*, pages 241–249, 1998. (Cité page 34.)

F. Liu et C. Yu. Personalized web search for improving retrieval effectiveness. *IEEE Transactions on knowledge and data engineering*, 16(1) :28–40, 2004. (Cité pages 9, 17, 18, 22, 24, 31, 43, 46, 70, 71 et 89.)

T. Su Louise. Evaluation measures for interactive information retrieval. *Information Processing and Management*, 28(4) :503–516, 1992. ISSN 0306-4573. (Cité page 87.)

C. Magennis et J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. Dans *Proceedings of the* 20^{th} *annual international ACM SIGIR conference on Research and development in information retrieval*, pages 324–332, 1997. (Cité page 20.)

R. Mandala, T. Takenobu, et T. Hozumi. Complementing wordnet with roget and corpus-based automatically constructed thesauri for information retrieval. Dans *Proceedings of the* 9^{th} *Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–101, 1999. (Cité page 49.)

J.P Mc Gowan. *A multiple model approach to personalised information access*. Master Thesis in computer science, Faculty of science, University College Dublin, 2003. (Cité pages 5, 17, 18 et 31.)

A. Micarelli, F. Gasparetti, F. Sciarrone, et S. Gauch. *Personalized search on the World Wide Web*. P. Brusilovsky and A. Kobsa and W. Nedjl, Springer Verlag Berlin Heidelberg, 2007. (Cité pages 5, 11, 12, 18, 29, 42 et 70.)

M. Mitra, A. Singhal, et C. Buckley. Improving automatic query expansion. Dans *Proceedings of the* 21^{st} *annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–214, 1998. (Cité page 49.)

S. Mizzaro. How many relevances in information retrieval ? *Interacting with Computers*, 10(3) :303–320, 1998. (Cité page 87.)

B. Mobasher. *Data mining for Web personalization*. Brusilovsky, P. Kobsa, A. Nejdl W., Lecture Notes in Computer Science. Springer Verlag, Berlin Heidelberg, NewYork, 2007. (Cité page 18.)

B. Mobasher, H. Dai, T. Luo, Y. Sun, et J. Zhu. Integrating web usage and content mining for more effective personalization. Dans *EC-Web*, pages 165–176, 2000. URL `http://citeseer.ist.psu.edu/mobasher00integrating.html`. (Cité pages 46 et 68.)

J. Mostafa, S. Mukhopadhyay, et M. Palakal. Simulation studies of different dimensions of users' interests and their impact on user modeling and information filtering. *Information Retrieval*, 6(2) :199–223, 2003. ISSN 1386-4564. (Cité page 88.)

F. Naumann, U. Leser, et J.C. Freytag. Quality-driven integration of heterogenous information systems. Dans *Proceedings of the* 11^{th} *Conference on Very Large Databases (VLDB)*, pages 447–458, 1999. (Cité page 32.)

R. Navarro-Prieto, M. Scaife, et Y. Rogers. Cognitive strategies in web searching. 2006. (Cité page 9.)

M. Nettleton, L. Calderon-Benavides, et R. Baeza-Yates. Analysis of web search engine query sesions. Dans *Text REtrieval Conference , Gaithersburg, Maryland, USA, 18/11/2003-21/11/2003*, page (electronic medium), December 2006. (Cité page 60.)

G. Paiattini, C. Calero, et A. Genero. *Information and database quality*. Kluwer Academics, 2006. (Cité page 32.)

G. Pasi. A logical formulation of the boolean model and of weighted boolean model. Dans *Proceedings of the Workshop on Logical and Uncertainty Models for Information Systems*, pages 1–11. ACM, 1999. ISBN 0-89791-674-3. (Cité page 20.)

M. J. Pazzani, J. Muramatsu, et D. Billsus. Syskill & webert : Identifying interesting web sites. Dans *Proceedings of the* 30^th *National Conference on Artificial Intelligence*, pages 54–61, Portland, 1996. (Cité pages 5, 31 et 68.)

J. Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann publishers Inc., San Francisco, CA, USA, 1988. ISBN 0-934613-73-7. (Cité page 80.)

D. Petrelli. On the role of user-centred evaluation in the advancement of interactive information retrieval. *Information Processing and Management*, 44(1) :22–38, 2008. ISSN 0306-4573. (Cité pages 89 et 105.)

K. Pinel-Sauvagnat et M. Boughanem. Using a relevance propagation method for Adhoc and Heterogeneous tracks in INEX 2004. Dans Nobert Fuhr, Mounia Lalmas, Saadia Malik, et Zoltan Szlavik, éditeurs, *INitiative for the Evaluation of XML Retrieval (INEX), Dagstuhl, Allemagne, 06/12/2004-08/12/2004*, LNCS, pages 337–348, http ://www.springerlink.com, décembre 2004. Springer. URL `http://www.irit.fr/~Karen.Pinel-Sauvagnat/fichiers/xfirm@inex04.pdf`. (Cité page 91.)

W. Pohl. *Logic-Based Representation and Inference for User Modeling Shell Systems*. PhD thesis, 1997. (Cité page 10.)

J. M. Ponte et W. B. Croft. A language modeling approach to information retrieval. Dans *Proceedings of the* 21^st *International ACM-SIGIR Conference*, pages 275–281. ACM, August 1998. (Cité page 48.)

M. P. Pryor. The effects of singular value decomposition on collaborative filtering. Rapport technique, Hanover, NH, USA, 1998. (Cité page 15.)

Y. Qui et H. P. Frei. Concept based query expansion. Dans *Proceedings of the* 16^th *annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 160–169, 1993. (Cité page 49.)

P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, et J. Riedl. GroupLens : An Open Architecture for Collaborative Filtering of Netnews. Dans *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pages 175–186, Chapel Hill, North Carolina, 1994. ACM.

URL citeseer.ist.psu.edu/resnick94grouplens.html. (Cité page 68.)

P. Resnik. Semantic classes and syntactic ambiguity. Dans *Proceedings of ARPA Workshop on Human Language Technology*, LNCS, pages 278–283. Association for Computational Linguistics (ACL), 1993. (Cité page 47.)

Hyoung R.K. et Philip K. Chan. Learning implicit user interest hierarchy for context in personalization. Dans *IUI '03 : Proceedings of the 8th international conference on Intelligent user interfaces*, pages 101–108, New York, NY, USA, 2003. ACM. (Cité page 18.)

S. E. Robertson et M. M. Hancock-Beaulieu. On the evaluation of ir systems. *Information Processing and Management*, 28(4) :457–466, 1992. ISSN 0306-4573. (Cité page 4.)

S.E. Robertson. The probability ranking principle in ir. *In Journal of documentation*, 33(4) :294–304, 1977. (Cité page 3.)

S.E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, et M. Lau. Okapi at trec 3. Dans *Text REtrieval Conference*, pages 21–30, 1992. (Cité page 5.)

S.E. Robertson, S. Walker, M. Sparck Jones, et al. Okapi at trec-3. Dans *Second Text Retrieval Conf (TREC-3)*, pages 109–26, 1995. (Cité page 93.)

J. Rocchio. Relevance feedback in information retrieval. Prentice Hall, 1971. (Cité pages 4, 20, 47 et 49.)

N. Ryan, J. Pascoe, et D.Morse. *Enhanced Reality Fieldwork : the Context-Aware Archaeological Assistant*. Gaffney, V., van Leusen, M., Exxon, S. (eds.) Computer Applications in Archeology, 1997. (Cité page 6.)

W. White Ryen, I. Ruthven, J. M. Jose, et C. J. Van Rijsbergen. Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems*, 23(3) :325–361, 2005. ISSN 1046-8188. (Cité pages 24 et 88.)

G. Salton. *Automatic information organisation and retrieval*. McGraw-Hill, New York, 1968. (Cité page 3.)

G. Salton, E. Fox, et H. Wu. Extended boolean information retrieval. *Communications of the ACM*, 26(11) :1022–1036, 1983. (Cité page 50.)

T. Saracevic. The stratefied model of information retrieval interaction : extension and applications. Dans *Proceedings of the 60th annual meeting of the American Society for Information Science. Medford, NJ*, pages 313–327, 1997. (Cité page 7.)

T. Saracevic, H. Mokros, L. Su, et A. Spink. Nature of interaction between users and intermdiaries in online searching. Dans *Williams, M. E Eds*, pages 329–341, 1991. (Cité page 5.)

B. N. Schilit, A. LaMarca1, G. Borriello, W. G. Griswold, D. McDonald, E. Lazowska, A. Balachandran, J. Hong, et V. Iverson. Ubiquitous location-aware computing and the place lab initiative challenge. Dans

The First ACM International Workshop on Wireless Mobile Applications and Services on WLAN (WMASH 2003), San Diego, CA, September 19, 2003, New York, NY, USA, 2003. ACM. (Cité page 9.)

B.N Schilit, N.I. Adams, et R. Want. Context-aware computing applications. Dans *Proceedings of the Workshop on Mobile Computing Systems and Applications*, pages 85–90. IEEE Computer Society, Santa Cruz, CA, 1994. (Cité page 6.)

H. Schutze et Pedersen J. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management (IPM)*, 33(3) :307–318, 1997. (Cité page 49.)

R.D Shachter. Probabilistic inference and influence diagrams. Dans *Operating Research*, volume 36, pages 589–604, 1988. (Cité page 77.)

L. Shamber. Relevance and information behaviour. *ARIST, William, M.E. Eds*, pages 3–48, 1994. (Cité page 4.)

X. Shen, B. Tan, et C. Zhai. Context-sensitive information retrieval using implicit feedback. Dans *Proceedings of the* 28^{th} *annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, New York, NY, USA, 2005. ACM. (Cité pages 9, 18, 24, 31, 46 et 47.)

L. Si et J. Callan. Unified utility maximization framework for resource selection. Dans *Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM)*, pages 32–41, 2004. (Cité page 32.)

A. Sieg, B. Mobasher, et R. Burke. User's information context : Integrating user profiles and concept hierarchies. Dans *Proceedings of the 2004 Meeting of the International Federation of Classification Societies*, numéro 1, pages 28–40, 2004a. (Cité page 9.)

A. Sieg, B. Mobasher, et R. Burke. Web search personalization with ontological user profiles. Dans *Proceedings of the* 16^{th} *ACM conference on Conference on information and knowledge management*, pages 525–534, New York, NY, USA, 2007. ACM. (Cité pages 17, 18, 22, 24, 31, 32, 43, 70, 71 et 88.)

A. Sieg, B. Mobasher, R. Burke, G. Prabu, et S. Lytinen. Using concept hierarchies to enhance user queries in web-based information retrieval. Dans *The IASTED International Conference on Artificial Intelligence and Applications. Innsbruck, Austria*, 2004b. (Cité pages 46 et 47.)

B. Smyth et E. Balfe. Anonymous personalization in collaborative web search. *Information retrieval*, 9(2) :165–190, 2006. (Cité page 9.)

M. Speretta et S. Gauch. Personalized search based on user search histories. Dans *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 622–628, 2005. (Cité pages 24, 32, 43, 46, 89, 95, 101 et 102.)

S. Sriram, X. Shen, et C. Zhai. A session-based search engine. Dans *Proceedings of the International ACM SIGIR Conference*, 2004. (Cité pages 31 et 43.)

L. Tamine. *Optimisation de requêtes dans un système de recherche d'information : approche basée sur les techniques avancées d'algorithmique génétique*. Thèse de Doctorat en Informatique, Université Paul Sabatier (Toulouse), 2000. (Cité page 64.)

L. Tamine et W. Bahsoun. Définition d'un profil multidimensionnel de l'utilisateur : vers une technique basée sur l'interaction entre dimensions. Dans *Conférence francophone en Recherche d'Information et Applications, Lyon, 15/03/2006-17/03/2006*, pages 225–236, http ://www.irit.fr/ARIA, mars 2006. Association Francophone de Recherche d'Information et Applications (ARIA). (Cité page 25.)

L. Tamine, F. Boubekeur, et M. Boughanem. On using graphical models for supporting context-aware information retrieval. Dans *International Conference on the Theory of Information Retrieval (ICTIR), Budapest (Hungary), 18/10/2007-20/10/2007*, pages 213–222, infoatinfota.org, octobre 2007a. Foundation for Information Society. (Cité pages 26 et 83.)

L. Tamine et M. Boughanem. Influence diagrams for contextual information retrieval. Dans *European Conference on Information Retrieval (ECIR), London, 10/04/2006-12/04/2006*, LNCS, pages 464–467, http ://www.springerlink.com, avril 2006. Springer. (Cité pages 26, 71 et 83.)

L. Tamine, M. Boughanem, et C. Chrisment. Accès personnalisé à l'information : vers un modèle basé sur les diagrammes d'influence. *Information - Interaction - Intelligence (I3)*, 6(1) :69–90, 2006a. (Cité pages 26 et 83.)

L. Tamine, M. Boughanem, et W.N. Zemirli. Inferring the user's interests using the search history. Dans Martin Schaaf et Klaus-Dieter Althoff, éditeurs, *Workshop on information retrieval, Learning, Knowledge and Adaptatbility (LWA), Hildesheim, Germany, 09/11/06-11/11/06*, pages 108–110, http ://www.uni-hildesheim.de, novembre 2006b. Institute of Computer Science. (Cité pages 25 et 33.)

L. Tamine, M. Boughanem, et W.N. Zemirli. Exploiting multi-evidence from multiple user's interests to personalizing information retrieval. Dans Youakim Badr, Richard Chbeir, et Pit Pichappan, éditeurs, *IEEE International Conference on Digital Information Management (ICDIM), Lyon, 28/10/07-30/10/07*, pages 7–12, http ://computer.org/cspress, octobre 2007b. IEEE Engineering Management Society. (Cité pages 26, 71, 77, 79, 81 et 83.)

L. Tamine, M. Boughanem, et W.N. Zemirli. Personalized document ranking : Exploiting evidence from multiple user interests for profiling and retrieval. *Journal of Digital Information Management, in press*, 2008a. (Cité pages 9, 71, 81 et 104.)

L. Tamine, C. Chrisment, et M. Boughanem. Multiple query evaluation based on an enhanced genetic algorithm. *Information Processing and Management*, 39(2) :215–231, 2003. ISSN 0306-4573. (Cité pages 5, 26, 54 et 63.)

L. Tamine, M. Daoud, B. Dinh, et M. Boughanem. Contextuel query classification on the web. Dans Martin Schaaf et Klaus-Dieter Althoff, éditeurs, *Workshop on information retrieval, Learning, Knowledge and Adaptatbility (LWA), Hildesheim, Germany, 09/11/08-11/11/08*, http ://www.unihildesheim.de, novembre 2008b. Institute of Computer Science. (Cité pages 26, 60, 61 et 64.)

L. Tamine, W.N. Zemirli, et W. Bahsoun. Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information. *Information - Interaction - Intelligence*, 7(1) :(en ligne), 2007c. URL `http://www.revue-i3.org/volume07/numero01/revue_i3_07_01_04.pdf`. (Cité pages 25 et 33.)

L. Tamine-Lechani, M. Boughanem, et M. Daoud. Evaluation of contextual information retrieval : overview of issues and research. *Knowledge and Information Systems, to appear*, 2009. (Cité pages 24, 87 et 104.)

L. Tamine-Lechani, M. Boughanem, et N. Zemirli. Personalized document ranking : exploiting evidence from multiple user interests for profiling and retrieval. to appear. *In Journal of Digital Information Management*, 2008. (Cité pages 25, 26, 27, 33, 44 et 83.)

B. Tan, X. Shen, et Ch. Zhai. Mining long-term search history to improve search accuracy. Dans *KDD '06 : Proceedings of the* 12^{th} *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 718–723, New York, NY, USA, 2006. ACM. ISBN 1-59593-339-5. (Cité page 31.)

Y. Tao, N. Mamoulis, et D. Papadias. Validity information retrieval for spatio-temporal queries. Dans *SSTD*, 2003. (Cité page 7.)

J. Teevan et S.T. Dumais. Personalizing search via automated analysis of interests and activities. Dans *Proceedings of the* 28^{th} *International SIGIR conference on research and development in information retrieval*, pages 449–456, 2005. (Cité pages 17, 18, 31, 46 et 70.)

M. Timothy, T. Sherry, et M. Robert. Hypermedia learning and prior knowledge : domain expertise vs. system expertise. *Journal of Computer Assisted Learning*, 21(12) :53–64, 2005. (Cité page 8.)

A. Tombros, I. Ruthven, et J.M. Jose. How users assess web pages for information seeking. *Journal of American Society of Information Science and Technology (JASIST)*, 56(4) :327–344, 2005. ISSN 1532-2882. (Cité page 9.)

P. Vakkari. A theory of the task-based information retrieval process : a summary and generalisation of a longitudinal study. *Journal of documentation*, 57(1) :44–60, 2001. (Cité page 5.)

C.J. van Rijsbergen. A non-classical logic for information retrieval. *The computer journal*, 29(6) :481–485, 1986a. (Cité page 3.)

C.J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29(6) :481–485, 1986b. (Cité page 48.)

Ellen M. Voorhees. Overview of trec 2001. Dans *TREC*, 2001. (Cité page 85.)

G.I. Webb, M. Pazzani, et D. Billsus. Machine learning for user modeling. *User Modeling and User-Adapted Interaction*, 11(1-2) :19–29, 2001. (Cité page 18.)

T. Westerveld, W. Kraaij, et D. Hiemstra. Retrieving web pages using content, links, urls and anchors. Dans *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, pages 663–672, 2001. (Cité pages 46 et 70.)

T. Westerveld, W. Kraaij, et D. Hiemstra. Retrieving web pages using content, links, urls and anchors. Dans Ellen M. Voorhees et Donna K. Harman, éditeurs, *Proceedings of the* 10^{th} *Text Retrieval Conference (TREC-10)*, pages 663–672. National Institute of Standards and Technology, NIST Special Publication 500-250, 2002. (Cité pages 21, 47 et 52.)

W.A. Woods. A better way to organize knowledge. Technical report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA, april 1997. (Cité page 47.)

H.I. Xie. Users' evaluation of digital libraries (dls) : Their uses, their criteria, and their assessment. *Information Processing and Management*, 44(3) : 1346–1373, 2008. ISSN 0306-4573. (Cité pages 9 et 18.)

J. Xu. *Solving the word mismatch problem through automatic text analysis*. Ph.D. Thesis, Department of Computer Science, University of Massachusetts, Amherst, MA, USA, May 1997, 1997. (Cité page 47.)

R. Yager. On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Transactions on Systems, Man and Cybernetics*, IEEE Transactions on Systems, Man and Cybernetics(1) :183–190, 1988. (Cité pages 21, 50 et 57.)

J.J Yang et R.R Korfhage. Query optimisation in information retrieval using genetic algorithms. Dans *Proceedings of International Conference on Genetic Algorithms*, pages 603–613, 1993. (Cité pages 53 et 62.)

S. Yau, L. Huan, D. Huang, et Y. Yao. Situation-aware personalized information retrieval for mobile internet. Dans *Proceedings of the* 27^{th} *Annual International Computer Software and Applications Conference (COMPSAC)*, 2003. (Cité page 46.)

L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning, parts i, ii. *Information Science*, 8 :199–249, 1975. (Cité page 51.)

W. N. Zemirli, L. Tamine, et M. Boughanem. Présentation et évaluation d'un modèle d'accès personnalisé à l'information basé sur les diagrammes d'influence. Dans *Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID), Perros-, 22/05/07-25/05/07*, pages 75–86, http ://inforsid.irit.fr, mai 2007. INFORSID. (Cité page 81.)

Titre De la recherche d'information orientée système vers la recherche d'information orientée contexte :
Verrous, contributions et perspectives

Résumé Le résumé en français (≈ 1000 caractères)

Mots-clés Les mots-clés en français

Title Le titre en anglais

Abstract Le résumé en anglais (≈ 1000 caractères)

Keywords Les mots-clés en anglais